MODELO DE FILTRADO COLABORATIVO DE LA CADENA DE MARKOV PARA LA RECOMENDACIÓN DE MATRÍCULA EN MATERIAS POR SEMESTRE ORIENTADO A LA REDUCCIÓN DEL TIEMPO MARGINAL UTILIZADO POR UN ESTUDIANTE PARA FINALIZAR SU CARRERA: CASO DE ESTUDIO, ISC EN LA UTP

TRABAJO DE GRADO COMO REQUISITO PARA OPTAR AL TÍTULO DE

MAGÍSTER EN ISC

FACULTAD DE INGENIERÍAS

POSTGRADOS

UTP

MAESTRÍA EN ISC

MODELO DE FILTRADO COLABORATIVO DE LA CADENA DE MARKOV PARA LA RECOMENDACIÓN DE MATRICULA EN MATERIAS POR SEMESTRE ORIENTADO A LA REDUCCIÓN DEL TIEMPO MARGINAL UTILIZADO POR UN ESTUDIANTE PARA FINALIZAR SU CARRERA: CASO DE ESTUDIO, ISC EN LA UTP

Juan Camilo Atehortua Zuluaga

Asesor

PhD. Jorge Iván Ríos Patiño

FACULTAD DE INGENIERÍAS

UTP

MAESTRÍA EN ISC

NOTA DE ACEPTACIÓN:

_____
_____
_____
_____
_____
_____
_____

_____
Nombre director, orientador, asesor

_____
Firma jurado (Nombres)

_____
Firma Jurado (Nombres)

Pereira - Risaralda, Mayo 2020

Dedico esta tesis a Dios por estar conmigo en todos los momentos apoyando y guiando mis proyectos, por poner a las personas adecuadas en mi vida, principalmente a mi esposa y a mis padres.

AGRADECIMIENTO

El autor expresa sus agradecimientos a:

Mi profesor PhD Jorge Iván Ríos; por guiarme de la manera más sabia para encontrar el camino cuando estaba perdido e impulsarme para superar cada obstáculo, por apoyarme para continuar en este proyecto profesional.

El ingeniero Carlos López por su ayuda sin importar el tiempo y horario, siempre dispuesto a enseñarme a ordenar mis ideas. Por la gestión que permitió el acceso a la información.

Mi esposa por ser el principal soporte en este proyecto, apoyando mis viajes de varios fines de semanas, por estar al pie del cañón en muchas de las desveladas, madrugadas y trasnochadas. Por ser mi apoyo incondicional y por la motivación que me brindo para seguir adelante.

La familia Bazurto Botero por brindarme un lugar caluroso, cómodo y amable donde llegar en la ciudad de Pereira.

TABLA DE CONTENIDO

Pág.

INTRODUCCIÓN....15
1 CAPITULO I: DEFINICIÓN DEL PROBLEMA....18
1.1. Descripción del Problema....18
1.2. Formulación del problema....20
1.3. Objetivos de la investigación....21
1.3.1 Objetivo general....21
1.3.2 Objetivos específicos....21
1.4. Justificación de la investigación....21
1.5. Viabilidad de la investigación....24
1.6. Alcance del estudio....25
1.7. Metodología....26
1.7.1 Hipótesis....26
1.7.2 Diseño....26
1.8. Sostenibilidad del Proyecto....30
2 CAPÍTULO II: APROXIMACIÓN AL ESTADO DEL ARTE....32
2.1. Inteligencia artificial....32
2.2. Marco teórico....33
3 CAPÍTULO III: MARCO CONCEPTUAL....56
3.1. Fundamentación teórica....56
3.1.1 Cadenas de Markov....56
3.1.2 Máximo estimado de verosimilitud....59

3.1.3 Sistema de colaboración de filtrado.....61
4 CAPÍTULO IV: DESARROLLO DE LA TESIS.....63
4.1. Nomenclatura.....63
4.2. Modelo matemático.....68
4.2.1 Modelo básico de Markov.....68
4.2.2 Modelo por omisión en cadenas de Markov.....70
4.3. Pseudocódigos.....72
4.3.1 Pseudocódigo para modelo básico de Markov.....72
4.3.2 Pseudocódigo para modelo por omisión en cadenas de Markov.....79
4.4. Resultados obtenidos.....82
4.5. Python.....84
5 CAPITULO V: REFERENCIA BIBLIOGRÁFICA, RECOMENDACIONES Y CONCLUSIONES.....86
5.1. Conclusiones y trabajos futuros.....86
5.2. Bibliografía.....88

ÍNDICE DE TABLAS

Tabla I. Variables del Modelo Formal.....63

Tabla II. Muestra de datos de inscripción de cuatro estudiantes en tres semestres consecutivos..70

Tabla III. Cadenas de dos y tres materias consecutivas matriculadas por un estudiante.....73

Tabla IV. Todas las cadenas de tres cursos consecutivos en el conjunto de datos iniciando con. 76

Tabla V: Materias recomendadas para cada estudiante.....79

Tabla VI: Cadenas de dos materias omitiendo el segundo semestre.....81

Tabla VII: Todas las cadenas omitidas de tres cursos consecutivos en el conjunto de datos iniciando con.....81

Tabla VIII: Materias recomendadas para el estudiante, en el modelo por omisión de cadenas de Markov.....82

Tabla IX: Rendimiento de modelo básico y por omisión en Cadenas de Marcov.....83

ÍNDICE DE GRÁFICAS

Fig. 1. Igualdad en listas circulares.....29

ÍNDICE DE ANEXOS

Anexo A … Materias matriculadas en semestres consecutivos por cuatro estudiantes

Anexo B ... Cadenas de dos materias consecutivas tomadas por un estudiante

Anexo C … Cadenas de tres materias consecutivas tomadas por un estudiante

Anexo D ... Todas las cadenas de tres cursos consecutivos en el conjunto de datos iniciando con $\{m_2, m_3\}$

Anexo E ... Materias recomendadas para matricular en el siguiente semestre a cada estudiante

Anexo F ... Materias recomendadas para matricular en el siguiente semestre a cada estudiante que no tuvo recomendación en el modelo básico de Markov, utilizando el modelo por omisión para la cadena de Markov

Anexo G … Archivo *admitidos_isc_graduados.json* el cual contiene la información de los estudiantes graduados en ISC

Anexo H … Archivo *notas_graduados_isc.json* el cual contiene información de materias con sus notas de los estudiantes graduados en ISC

Anexo I … Archivo *get_admitidos_isc_graduados.py* con el código fuente para obtener la información de los graduados en ISC

Anexo J … Archivo *get_notas_graduados_isc.py* con el código fuente para obtener las materias con sus respectivas notas de los graduados en ISC

Anexo K … Archivo *new_column_semester_in_notas_graduados.py* con el código fuente para calcular el semestre en el que fue cursada la materia

Anexo L … Archivo *chains_two_three_consecutive_matter.py* con el código fuente con el modelo básico de cadenas de Markov y por omisión

RESUMEN

El rendimiento académico en el programa de ISC de la UTP ha venido en decremento debido al aumento de ingreso de estudiantes a la educación superior, estos finalizan los estudios al menos en dos semestres más que los estándares utilizados por el MEN.

En cuanto a la solución del problema se propone elaborar un modelo de filtrado colaborativo de Cadena Básica de Markov y otro modelo de Cadena de Markov por Omisión, el cual permite formular y resolver por medio del comportamiento estocástico de las materias que ha tomado un estudiante en semestres anteriores y con esta información realizar la recomendación de matrícula en materias por semestre orientado a la reducción del tiempo marginal utilizado por el estudiante para finalizar su carrera.

Se realizó un análisis para cuatro estudiantes del conjunto de prueba en el que se calculó el Recall como, el porcentaje de materias que se recomendaron por el sistema para matricular en el siguiente semestre y que ya fueron tomadas por los estudiantes en sus semestres anteriores. Por otra parte, se calculó la precisión como, el porcentaje de materias que fueron matriculadas en el siguiente semestre basado en los datos de prueba.

Palabras clave: Inteligencia artificial, Cadena de Markov, Filtrado colaborativo, Sistemas de recomendación, Máximo estimado de verosimilitud, Reducción del tiempo marginal, Semestres académicos.

ABSTRACT

The academic performance in the program systems and computer engineering of the UTP has been declining due to the increase in the entrance of students to higher education, they finish their studies in at least two semesters more than the standards used by the MEN.

Regarding the solution of the problem, it is proposed to develop a collaborative filtering model of Markov Basic Chain and another model of Markov Chain by Omission, which allows formulating and solving through the stochastic behavior of the subjects that a student has taken in previous semesters and with this information make the recommendation of enrollment in subjects per semester aimed at reducing the marginal time used by the student to finish his career.

An analysis was carried out for four students of the test set in which the Recall was calculated as, the percentage of subjects that were recommended by the system to enroll in the following semester and that were already taken by the students in their previous semesters. On the other hand, the precision was calculated as, the percentage of subjects that were enrolled in the following semester based on the test data.

Keywords: Artificial intelligence, Markov chain, Collaborative filtering, Recommendation systems, Maximum likelihood estimates, Marginal time reduction, Academic semesters.

GLOSARIO

UTP: Universidad Tecnológica de Pereira

ISC: Nombre abreviado que permite identificar rápidamente al programa de Ingeniería de Sistemas y Computación en la UTP.

OECD: Organización para la Cooperación y el Desarrollo Económicos

SEDLAC: Centro de Estudios Distributivos, Laborales y Sociales

MEN: Ministerio de Educación Nacional de Colombia

MLE: Máximo Estimado de Verosimilitud

UNESCO: La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura

IESALC: El Instituto Internacional de la UNESCO para la Educación Superior en América Latina y el Caribe

INTRODUCCIÓN

El tiempo marginal que un estudiante de ISC toma después de los 10 semestres académicos para finalizar sus estudios universitarios ha estado en aumento desde el año 2014 y junto con el incremento de ingresos de estudiantes a la educación superior para los años 2014 al 2019 en el programa de ISC se ha aumentado la deserción para estos mismos periodos [1].

Con esto se quiere decir que el éxito en los estudios universitarios se basa en dos aspectos: la duración de los estudios y la probabilidad de graduarse realmente en uno de los programas académicos [2]. Es de aclarar que aunque es un tema realmente importante para la educación en la UTP, actualmente no se encuentra un sistema que utilice inteligencia artificial o sistemas de recomendación para estos procesos.

En cuanto al histórico de datos, el espacio muestral incluye todo el conjunto de todos los estudiantes de la UTP, en una base de datos que en términos numéricos o lingüísticos contiene información académica, demográfica y socio-económica. Para nuestro espacio muestral se tomó un subconjunto de interés que cumple con las condiciones de ser un estudiante de ISC que inicio en una cohorte en el 2010 y hasta el primer semestre de 2019 ya se graduó en su carrera universitaria para un total de 977 estudiantes matriculados en todo este tiempo.

Con respecto a la solución del problema, se utilizó la cadena de Markov más simple en donde una secuencia de materias matriculadas por estudiantes son representadas en un proceso estocástico y el conjunto de probabilidades de transición que son calculadas desde los datos con la probabilidad condicional para entrar en cada estado, dado el estado inmediatamente anterior [3]. Se utilizaron dos enfoques para estimar las probabilidades de transición:

1. Modelo básico con MLE
2. Modelo de estimación mejorado con MLE basado en un modelo por omisión de semestres.

Prosiguiendo con el análisis, existen diferentes sistemas de recomendación con diferentes técnicas, este proyecto se centra en uno de los métodos tradicionales como lo es el método basado en el filtrado colaborativo. Las técnicas de recomendación basadas en el filtrado colaborativo ayudan a las personas a tomar decisiones basadas en las opiniones de otras personas que comparten intereses similares [4]. Se planteó un modelo colaborativo de filtrado con cadenas de Markov que permita la reducción de los semestres que un estudiante toma para finalizar su carrera con éxito por medio de la recomendación de las materias que un estudiante debe tomar en el siguiente semestre.

Este estudio consta de cinco grandes capítulos, en el capítulo I se introduce al problema con su respectivo alcance y metodología utilizada en la investigación.

En el capítulo II y III se explica la aproximación al estado del arte y el marco teórico donde se investigó sobre los diferentes modelos científicos utilizados para disminuir el tiempo marginal que un estudiante toma para finalizar su carrera. Se encuentran algunos artículos que se apoyan en conocimientos de inteligencia artificial utilizando datos socio-demográficas y académicas para mejorar el rendimiento académico como género, estatus económico, tipo de escuela y desempeño [5][6]. A pesar que estos estudios se basan en los datos anteriores, en este estudio se definió utilizar las variables académicas que intervienen en el tiempo marginal como las notas de cada una de las materias por semestre utilizando el histórico de datos de materias que matricularon los estudiantes.

Otro capítulo importante es el Capítulo IV en donde se especifica el espacio muestral utilizado y que procedimientos se realizaron en los datos, además de esto, se encuentra el modelo matemático utilizado en donde se aplica un modelo de Cadenas de Markov orientado al progreso de cohortes de estudiantes, pero utilizado en específico a la manera como progresan los estudiantes a través de grupos de materias por semestre que hacen parte de un dominio especifico en el proceso de formación profesional [7]. Además se implementaron los temas relacionados con los pseudocódigos y los resultados obtenidos en las pruebas realizadas en donde se describen los pasos implementados del modelo de Markov para predecir las materias que cada estudiante es más probable que tome en su siguiente semestre.

Y en el último capitulo se encuentran las conclusiones y bibliografía.

1 CAPITULO I: DEFINICIÓN DEL PROBLEMA

1.1. Descripción del Problema

Para La Organización IESALC – UNESCO: “La deserción contribuye a generar inequidad y desequilibrios sociales y desvirtúa los objetivos que la sociedad le ha entregado a la educación superior“. En tal sentido, la composición de diferentes factores sociales, económicos, familiares e individuales afectan directamente al rendimiento académico del estudiante e incluso pueden llevar a la decisión de abandono de los estudiantes de la educación superior.

Las estadísticas de deserción indican que Colombia se encuentra un 17,8% por encima de la media de la deserción en educación superior. El 31% es la media alcanzada en los países correspondientes a la OECD y Colombia con un 48,8% debe aportar al mejoramiento en políticas en educación superior específicamente en calidad, pertinencia, y cobertura [8]. Entre los países latinoamericanos, Colombia se encuentra entre los que más deserción tuvieron en el año 2016 junto a Bolivia y Nicaragua según el SEDLAC [9].

Desde otra perspectiva, el rendimiento académico tiene un 24,83% del total de estudiantes a nivel nacional que se gradúan a tiempo; los demás estudiantes finalizan los estudios al menos en dos semestres más que los estándares utilizados por el Ministerio de Educación MEN (10 periodos académicos para estudios universitarios) [10]. Para el ministerio de educación la proporción de estudiantes que son matriculados dos semestres atrás son clasificados como desertores un año después.

Aunado a esto, el aumento de ingreso de estudiantes a la educación superior para el año 2010 a 2019 con 2824 estudiantes, específicamente en la UTP para el programa de ISC en la jornada diurna, se presentó un aumento de 783 estudiantes matriculados y esto a llevado a un incremento en la deserción entre los mismos periodos del 22,34%. Y para la jornada nocturna se presentó un número de matriculados de 320 estudiantes en el mismo periodo mencionado, en donde desertaron o cambiaron de programa un 54,63% de estudiantes [1].

En la UTP para el programa de ISC en jornada nocturna, ha venido presentando la tasa de deserción estudiantil más alta de la universidad, entre el año 2007 y 2019, en la mayoría de sus estadísticas supera el 20% de estudiantes que desertan, por otra parte la ISC (diurna), se encuentra con una deserción normal, localizándose con un valor menor a 10% en la mayoría de los años analizados. Para el año 2016 se obtuvo un 8,72% de deserción estudiantil comparado con un 21,28% para la jornada especial [1]. Lo que indica que hay una necesidad de desarrollar un método para optimizar el tiempo que toma un estudiante para finalizar su carrera, y que permita reconocer cuales son los factores más importantes que le permiten a la persona continuar con sus estudios.

Un estudio realizado para el programa de Ingeniería en Sistemas y Computación de la UTP indica que el retraso o deserción de los estudiantes en la educación superior depende de factores académicos, individuales, socioeconómicos e institucionales. Para este estudio se encontró que en la mayoría de los estudiantes en los primeros cuatro semestres tienen problemas con asignaturas como: matemáticas, física, programación y álgebra lineal. Además de estos factures académicos se encontró que el 29, 8% desertan por problemas económicos y un 26% de estos estudiantes cambian de carrera [11].

1.2. Formulación del problema

Los estudios académicos para un estudiante están basados en dos aspectos: la duración de sus estudios y la probabilidad de graduarse realmente en uno de los programas académicos. El aumento del tiempo de un estudiante en su fase de formación académica, siendo este el tiempo marginal que el estudiante toma por encima de diez semestres para la finalización de su carrera universitaria, tiene como causa principal que el estudiante suspenda sus estudios después de tener un número de créditos avanzados, este fenómeno conlleva a la pérdida de confianza y en el peor de los casos el estudiante se retira de la universidad aumentando directamente la carga de trabajo académico y gastos universitarios [2].

Teniendo en cuenta las cifras expresadas para el problema, tanto de deserción como de graduación a tiempo, cuando una persona se retira del sistema de educación superior o retrasa su finalización, específicamente en la carrera de ISC, está cerrando puertas no solo a la herramienta que le permite el crecimiento constante, sino, a las oportunidades que le permitan integrarse en un ambiente laboral con mejores oportunidades de desarrollar sus capacidades intelectuales y destacarse para continuar avanzando y aportando a los diferentes aspectos que permiten el crecimiento del sistema académico, económico y social. Es por esto, que se requiere mejorar el rendimiento académico, revisar sus causas, sus consecuencias, e implementar un modelo que permita reconocerlo y medirlo para brindar un proceso educativo optimizado a los estudiantes del programa de ingeniería en sistemas y computación.

1.3. Objetivos de la investigación

1.3.1 Objetivo general.

Construir un modelo formal, empleando filtrado colaborativo de la cadena de Markov para la reducción del tiempo marginal utilizado por un estudiante para la finalización de su carrera: caso de estudio, ISC en la UTP.

1.3.2 Objetivos específicos.

- Definir un conjunto de variables y elementos que intervienen en el tiempo marginal que gasta un estudiante en la terminación de su carrera de ingeniería, concretamente para nuestro caso de estudio ISC en la UTP.
- Investigar métodos formales que permitan la reducción del tiempo marginal de un estudiante para la terminación de su carrera, caso de estudio ISC en la UTP.
- Plantear un modelo formal que represente el tiempo marginal que un estudiante se gasta en la terminación de su carrera de ISC.
- Implementar un modelo computacional que permita la reducción de los semestres que un estudiante toma para finalizar su carrera con éxito.

1.4. Justificación de la investigación

De acuerdo al análisis realizado para la SPADIES entre los periodos de 2018-1 a 2019-1, la UTP tiene una tasa de deserción interanual por programa de 9,07% considerando como desertor al

estudiante que cambia de programa. Para el programa de ISC 40 estudiantes desertan, 22 cambian de programa, 91 se gradúan y 606 continúan en el programa para un total de 759 estudiantes, lo que indica que el 8,17% están desertando o cambian de programa y para el programa de ISC (Nocturno) 28 estudiantes desertan, 3 cambian de programa, 18 se gradúan y 228 continúan en el programa para un total de 759 estudiantes, lo que indica que el 11,19% están desertando o cambian de programa [12]. A pesar de que las estadísticas muestran un número de estudiantes que continúan en el programa, es probable que estas personas en los próximos años pueden desertan, por lo tanto es pertinente en el ámbito tecnológico utilizar las herramientas de Inteligencia artificial que permitan optimizar el rendimiento académico de estos estudiantes para evitar su deserción y permitir que el número máximo de semestres para su graduación sea entre 10 y 12.

Por otra parte, un estudio realizado entre el 2011 y 2014 señala que en Pereira y su área metropolitana hay un total de 42,826 estudiantes, el 87% en Pereira con un total de 37.246 estudiantes [13]. Se evidencia un crecimiento del 41% para el número de estudiantes que ingresan a la educación superior, el cual paso de 26.697 en el 2010 a 45.904 en 2017, este comportamiento de aumento se encuentra principalmente entre el año 2011 y 2014, luego de este año se encuentra un crecimiento de la población universitaria de 6% entre el año 2014 y 2017. Se encontró que casi la mitad de los estudiantes de pregrado con un 46% tienen procedencia de municipios que no hacen parte del área metropolitana, estas personas viven en la ciudad de Pereira en casa de familiares o conocidos y aproximadamente uno de cada tres alterna sus estudios con actividades laborales [9], hay que mencionar además, que el 61% de estos estudiantes van a una universidad pública [13].

Para el año 2019 ingresaron 8056 estudiantes a la UTP, donde 585 estudiantes se matricularon en ISC y 135 en ISC (Diurno) [15]. Considerando el crecimiento de la población para la educación superior y el problema que se genera a través del bajo rendimiento académico, generando altos índices de deserción, es relevante para el ámbito social y humanístico elaborar un modelo estocástico que permita determinar el rendimiento de los estudiantes a lo largo de un plan de estudios, para este caso, ISC, en la UTP. Es necesario recalcar que las investigaciones y resultados de este modelo serian importantes para el desarrollo continuo de la calidad en las políticas académicas y las estrategias de planificación de la Universidad.

El programa de ISC de la UTP es un programa acreditado de alta calidad por medio de la Resolución N° 16816 / Ago 19 – 2016 con Vigencia de 4 años [16], lo cual le permite tener un proceso de fomento, reconocimiento y mejoramiento continuo de la calidad. Los anteriores conceptos se conservaran por medio de un modelo que permita optimizar el rendimiento académico y disminuir los índices de deserción estudiantil en el programa de ISC que aportan en el ámbito económico para la universidad, y así mismo que permita mejorar los siguientes aspectos importantes para la calidad académica:

- Aumento del número de graduados desde que inicia una cohorte hasta finalizar sus estudios en 10 semestres y a un máximo de 12 semestres.
- Estabilidad de los ingresos transferidos a la Universidad, minimizando el costo por estudiante.
- Mejoramiento de índices de gestión para el apoyo a la calidad académica.
- Aumento de profesionales con mayor probabilidad de crecimiento económico y social para la ciudad de Pereira.

- Avance en conocimiento tecnológico el cual le permite a las empresas avanzar de una manera rápida.

Por otra parte, este proyecto no tiene aporte y no está buscando favorecer en cuanto a temas relacionados con el medio ambiente.

1.5. Viabilidad de la investigación

En cuanto a la solución del problema se propone elaborar el modelo de filtrado colaborativo de la cadena de Markov, el cual permite formular y resolver el problema con nuestros parámetros de incertidumbre o inciertos para cada uno de los estudiantes, como por ejemplo el comportamiento estocástico de las materias que ha tomado en semestres anteriores y las que tomara en los siguientes semestres. Es necesario usar métodos que consideren esas incertidumbres dentro del proceso de identificación de la mejor estrategia académica que debe ser tomada para resolver el problema del rendimiento académico. Para formular y resolver el problema se cuenta con equipos de cómputo y las personas con conocimientos en el área de modelos estocásticos.

Tomando en consideración el objetivo de esta tesis, se busca resolver el problema planteado y descrito con un modelo formal, en el cual las decisiones de localización y capacidad son hechas en el primer semestre, teniendo la posibilidad de revisar la decisión después del primer año cumpliendo con los requerimientos establecidos por el programa de ISC para la calidad académica. Teniendo en consideración lo anterior se cuenta con el histórico de datos necesario para realizar el análisis.

Hay que mencionar que, es un proyecto que requiere en su parte inicial el acceso a la información, la cual ya se ha brindado por parte de la Universidad Tecnológica de Pereira al equipo de investigación; además que se tiene la capacidad tecnológica en cuanto a herramientas y conocimientos del investigador, director y ayuda externa para obtener los datos, preprocesarlos y realizar el respectivo análisis que permita encontrar la solución al problema planteado inicialmente.

Además de esto, es necesario tener un computador para realizar las operaciones y finalmente el conocimiento del investigador, director y ayuda externa para el acceso y preprocesamiento de la información. Se debe adicionar un costo relacionado con los viajes y viáticos del investigador, el cual estará asistiendo a clases magistrales de procesos estocásticos.

Ahora bien el tiempo requerido para la solución a este problema es de máximo doce meses teniendo en cuenta que el tiempo destinado es de 7,5 horas por semana que en total suman 400 horas para la finalización de esta tesis. No solo el tiempo es considerado sino también se requiere de un experto en el área de la educación e inteligencia artificial; estas personas se encuentran en la ciudad de Pereira, por lo que se requiere un recurso económico para los viáticos y el viaje a esta ciudad, lo cual se calcula en máximo 25 viajes para los 12 meses de finalización para un total de un 3.750.000 de pesos.

1.6. Alcance del estudio

Con el propósito de cumplir con los objetivos propuestos, se implementó un modelo de filtrado colaborativo de Markov que recomienda al estudiante las materias que debe matricular en el

siguiente semestre, este proceso se pudo realizar gracias al acceso a los datos que contienen las variables académicas que intervienen en el tiempo marginal, como las notas de cada una de las materias por semestre en el histórico de datos de materias que matricularon los estudiantes, para la finalización de su carrera en ISC en la UTP.

1.7. Metodología

1.7.1 Hipótesis

El filtrado colaborativo de la cadena de Markov permite la generación de estrategias para la reducción del tiempo marginal utilizado por un estudiante para la finalización de su carrera: caso de estudio, ISC en la UTP.

1.7.2 Diseño

Con respecto a lo que se planteó en el problema de esta investigación, se realizaron una serie de procesos para proponer un avance de los procedimientos educativos en la decisión que un estudiante debe tomar para matricular unas materias que le permitan avanzar de una manera eficiente y óptima en el programa de ISC en la UTP, en efecto se realizó un estudio de tipo experimental con un enfoque cuantitativo [17]. Con el propósito de cumplir con los objetivos propuestos en esta investigación se tuvo un desarrollo metodológico basado en los siguientes ciclos:

1) En el ciclo inicial se definió usar las variables académicas que intervienen en el tiempo marginal como las notas de cada una de las materias por semestre utilizando el histórico

de datos de materias que matricularon los estudiantes para la finalización de su carrera en ISC en la UTP.

2) En el segundo ciclo se realizó una investigación que permitiera obtener el conocimiento que se encuentra en el conjunto de datos histórico para la matrícula realizada en cada semestre por cada estudiante, como resultado de este proceso se obtuvo que las cadenas de Markov permiten utilizar la transición de materias en cada semestre utilizando ese valor importante de cual materia matriculo primero para continuar con materias posteriores en los siguientes semestres, hay que decir También que se utilizó un método para ajustar modelo y estimar el puntaje de recomendación que tiene una materia que posiblemente debería tomar el estudiante en el siguiente semestre.

3) Se planteó un modelo colaborativo de filtrado con cadenas de Markov que permita la reducción de los semestres que un estudiante toma para finalizar su carrera con éxito por medio de la recomendación de las materias que un estudiante debe tomar en el siguiente semestre. Con respecto a la implementación del modelo matemático se diseñó un algoritmo en el lenguaje de Python para realizar las pruebas con cuatro estudiantes que ya están graduados que permitieran afinar y reconocer la precisión y el Recall. De estos cuatro estudiantes se tiene la información necesaria como notas por materia matriculadas en cada semestre.

El espacio muestral incluye todo el conjunto de todos los estudiantes de la UTP, es una base de datos que en términos numéricos o lingüísticos contiene información académica, demográfica y socio-económica. De nuestro espacio muestral se tomó un subconjunto de interés que cumple con las condiciones de ser un estudiante de ingeniería en sistemas y computación que inicio en

una cohorte en el 2010 y hasta el primer semestre de 2019 ya se graduó en su carrera universitaria para un total de 977 estudiantes matriculados en todo este tiempo.

Antes de realizar el proceso de recomendación hay que tener unos datos limpios en el que se apliquen procedimientos de reducción, de re-codificación, de integración y de una respectiva limpieza respectiva de los mismos.

Al hacer un análisis de los datos se observa que solo a partir del año 2010 se tienen los datos de algunos semestres, esto indica que si un estudiante inicio en el año 2008, no se tendría información de los primeros cuatro semestres.

Se realizó un ajuste a los datos indicando el número de semestre en que tomó cada materia, este se ejecutó con el siguiente algoritmo:

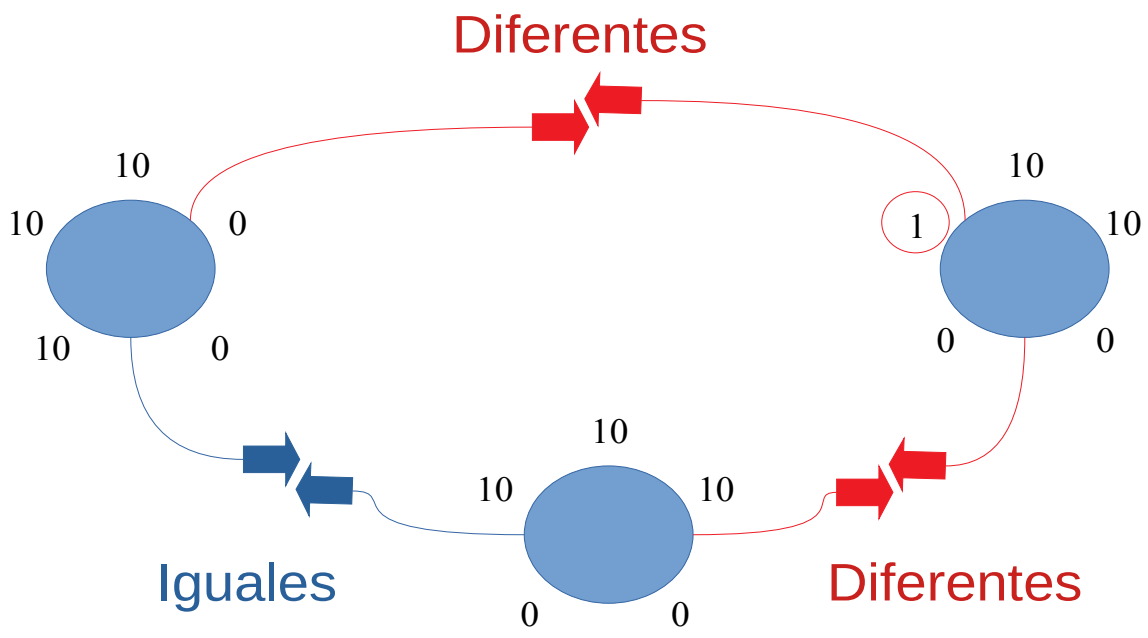
1 **Inicia** /*Algoritmo Preprocesamiento de semestres */
2 **Ordenar** las materias por año y semestre
3 **Seleccionar** el año en que se tomó la materia
4 **Seleccionar** año de admisión del estudiante
5 **Seleccionar** el semestre en que tomó la materia
6 **Calcular** El número de semestre**.**
((año_nota_materia – año_admisión) * 2) + (sem_nota_materia – sem_admisión
7 **Si** número de semestre es diferente de uno y diferencia de semestres mayor a uno **Entonces**
8 número de semestre := ((Año en que tomó la materia – año de admisión) * 2) + (Semestre en que tomó la materia – semestre de admisión)
Si no
Número de semestre := 1
9 **Termina**

En el punto cinco se inicia desde el año de admisión, se compara el año en la que inicio el primer semestre con el año de la nota de la materia, si es primer semestre y la diferencia de este cálculo indica que es un año diferente se restan los años y se multiplican por dos ya que cada año tiene

dos semestres; luego se hace una resta de los semestres que indica si hay que sumar o restar semestres dependiendo del semestre de admisión y el semestre de la nota en la materia.

Por otra parte, en el punto siete se recomienda comparar los conjuntos de materias de los estudiantes para la recomendación de materias que debería matricular en el siguiente semestre, se desarrolló en Python la técnica para validar si dos listas son idénticamente circulares. Se obtienen dos listas y se valida si hay identidad circular o no, como se muestra en la Figura 1.

Fig. 1. Igualdad en listas circulares. [18].



Y finalmente para la recomendación se creó el siguiente algoritmo:

1 **Inicia** /*Algoritmo para listas de materias */
2 **Seleccionar** los estudiantes con las notas y su respectivo semestre
3 **Agrupar** las materias de cada estudiante
4 **Ordenar** las notas de las materias por semestre de cada estudiante
5 **Adicionar** las materias en una lista
6 **Termina**

En el quinto punto se insertan las materias del primer semestre con el segundo, del segundo con el tercero, así sucesivamente. Proveer conjuntos de 2 o más materias, esto depende del semestre en el que se encuentre el estudiante.

1.8. Sostenibilidad del Proyecto

Este proyecto nace de la necesidad de resolver el problema del tiempo marginal que toma un estudiante para finalizar sus estudios académicos. Por lo tanto, para que este proyecto sea llevado a ejecución se plantea un estudio de viabilidad desde diferentes puntos como impacto social, económico, tecnológico y humanístico; es importante adicionar que este proyecto no aporta en un impacto ambiental.

Para iniciar con este proyecto se centró en el impacto social y humanístico, y se evaluó que existe la necesidad de una solución que permita disminuir el tiempo marginal que un estudiante toma para finalizar sus estudios académicos, esto basado en el estudio realizado en el proyecto de grado *"Nivel de deserción universitaria en el programa de ingeniería en sistemas y computación de la UTP"[1],* abriendo puertas para el crecimiento constante y así tener más oportunidades que le permitan integrarse en un ambiente laboral con mejores oportunidades de desarrollar sus

capacidades intelectuales y destacarse para continuar avanzando y aportando a los diferentes aspectos que permiten el crecimiento del sistema académico, económico y social.

Por otra parte, cabe mencionar el impacto económico del proyecto para el programa de ISC en la UTP, el cual le brindará ahorrar tiempo, dinero y espacios universitarios con un proceso educativo optimizado para los estudiantes, permitiendo a estos finalizar más rápido su carrera universitaria y así ingresar a la vida laboral para su estabilidad económica y profesional.

Este proyecto se basó sobre la actividad tecnológica para identificar distintos focos de investigación, tales como el impacto tecnológico en la educación para mejorar las condiciones de educación y abrir nuevas vías a la construcción de planes de estudios óptimos recomendados por el sistema a través de un histórico de estudiantes que hayan logrado graduarse en tiempos óptimos. Esta investigación aplica un avance en inteligencia artificial y recomendación automática, permite a su vez grandes innovaciones al sistema educativo que mejoran el tiempo de espera para que los estudiantes finalicen sus carreras universitarias. Además de permitir a la UTP beneficiarse de investigaciones tecnológicas que le ayudan a mejorar la calidad de sus programas y permitir mejorar el futuro de sus estudiantes.

2 CAPÍTULO II: APROXIMACIÓN AL ESTADO DEL ARTE

2.1. Inteligencia artificial

La inteligencia artificial (IA), es un término acuñado por primera vez a John McCarthy en 1956 [19], y la definió como: "… la ciencia e ingenio de hacer máquinas inteligentes, especialmente programas de cómputo inteligentes". Es una rama de la informática que se centra en el desarrollo de sistemas informáticos para simular los procesos de resolución de problemas y para duplicar las funciones del cerebro humano y se utilizan para resolver problemas inherentes en sistemas complejos. Por lo tanto, la IA puede definirse como los métodos utilizados para resolver problemas complejos basados en el comportamiento inteligente de los humanos y otras formas de vida animadas [20].

Informalmente hablando, el término inteligencia artificial se aplica cuando una máquina imita las funciones «cognitivas» que los humanos asocian con otras mentes humanas, como: "aprender"; y "resolver problemas". Por ejemplo, el reconocimiento óptico de caracteres ya no se percibe como un ejemplo de la "inteligencia artificial" habiéndose convertido en una tecnología común [21].

Otras observaciones de la IA afirman que es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos en base a dos de sus características primordiales: el razonamiento y la conducta [22].

Para Nils John Nilsson son cuatro los pilares básicos en los que se apoya la inteligencia artificial [19]:

- Búsqueda del estado requerido en el conjunto de los estados producidos por las acciones posibles.
- Algoritmos genéticos (análogo al proceso de evolución de las cadenas de ADN).
- Redes neuronales artificiales (análogo al funcionamiento físico del cerebro de animales y
- humanos).
- Razonamiento mediante una lógica formal análoga al pensamiento abstracto humano.

También existen distintos tipos de percepciones y acciones, que pueden ser obtenidas y producidas, respectivamente, por sensores físicos y sensores mecánicos en máquinas, pulsos eléctricos u ópticos en computadoras, tanto como por entradas y salidas de bits de un software y su entorno.

Varios ejemplos se encuentran en el área de control de sistemas, planificación automática, la habilidad de responder a diagnósticos y a consultas de los consumidores, reconocimiento de escritura, reconocimiento del habla y reconocimiento de patrones. Los sistemas de IA actualmente son parte de la rutina en campos como economía, medicina, ingeniería y la milicia, y se ha usado en gran variedad de aplicaciones de software, juegos de estrategia, como ajedrez de computador, y en la educación [23].

2.2. Marco teórico

Para este estudio se determinaron las proporciones de estudiantes que avanzan en cada materia, por cohorte. Estas proporciones son estimaciones de las probabilidades de avanzar, de acuerdo

con un enfoque frecuencial. Se hizo la validación de los supuestos necesarios para conformar el modelo Markoviano [24].

"Artificial Intelligence Assisted Higher Education: Early recognition System with Machine Learning to support Students Success", es un artículo que nos demuestra como la educación y los medios para educar han ido y siguen evolucionando para unificar la tecnología con el arte de enseñar, promoviendo la educación 4.0 [25].

Este artículo enfatiza acerca de la evolución tecnológica que va teniendo la educación. Es importante resaltar que la práctica docente también ha ido evolucionando ya que en un principio la enseñanza estaba basada netamente en el docente como una persona que solo daba clases magistrales; sin embargo este concepto ha ido cambiando ya que las diferentes corrientes pedagógicas han tomado al estudiante como uno de esos ejes centrales de conocimiento y es desde allí que la educación 4.0 va adquiriendo poder, bien lo dice el artículo cuando pronostica que en el 2020 se usarán más herramientas tecnológicas basadas en la inteligencia artificial.

Apoyar a los estudiantes con dispositivos portátiles y sensores inteligentes es una de las principales preocupaciones de la futura Educación 4.0. International Data Corporation (IDC) pronostica que el mercado general de auriculares con Realidad Aumentada y Virtual creció a 8.9 millones de unidades en 2018. Ese crecimiento continuará durante todo el período de pronóstico, llegando a 65.9 millones de unidades para 2022.

Este cambio pedagógico de resaltar y poner al estudiante como centro del aprendizaje no solo lo podemos encontrar en las instituciones de bachillerato, sino también en la universidad que por décadas su eje central son los estudiantes, ahora bien si lo ligamos a la parte tecnológica podríamos partir diciendo que la autorregulación del aprendizaje académico es el control que los

estudiantes tienen sobre su cognición, comportamiento, emociones y motivación mediante el uso de estrategias personales para archivar las metas que han establecido.

El modelo de Zimmerman de aprendizaje autorregulado es un modelo cíclico muy completo. Presenta información sobre tres fases (previsión, rendimiento y autorreflexión) y una variedad de métodos de evaluación de aprendizaje establecidos.

Para motivar, especialmente a los estudiantes de primer año, se mostró un diagrama de actividades de antiguos estudiantes. Esto subraya las actividades en el curso de matemáticas durante el semestre y las calificaciones / puntajes (puntajes altos de quienes aprueban el examen) dependiendo de las actividades completadas. Para ofrecer un acceso personalizado a los objetos de aprendizaje (LO), se necesitó un modelo de estudiante.

Se utilizó este modelo para encontrar enfoques que permitieron motivar y personalizar el modelo del estudiante. Apoyar a los estudiantes a autorregularse en términos de habilidades para la vida, y no solo en el entorno universitario, para que los estudiantes sigan aprendiendo de forma práctica y continua fuera del aula (aprendizaje de larga duración).

En 1950, Skinner, inaugura la primera máquina de enseñanza positivamente sobre los objetivos de aprendizaje, basado en estímulos, respuestas y recompensas de lo correcto En los próximos años un aspecto crucial es considerar el efecto de la emoción, la motivación y el efecto del ambiente de trabajo [21].

Education 4.0 utiliza métodos de inteligencia artificial, en cursos en línea, incluida una presencia interactiva en forma de aprendizaje combinado y características impulsadas por inteligencia artificial: un proceso de aprendizaje personalizado, aprendizaje práctico con realidad virtual

(VR) / realidad aumentada (AR) libro interactivo, modelo Skinner y video interactivo, modelo N.A. Crowder.

Los primeros reconocimientos en sistemas con Machine Learning se identifican en el curso de Gestión del Conocimiento con 38 estudiantes en riesgo en Administración de Empresas y 20 estudiantes en riesgo en Informática Empresarial, Donde se mejoró la tasa de fracaso en exámenes con casi la mitad. La idea contribuye fuertemente con el rendimiento de los alumnos.

Un segundo antecedente para basar esta tesis es *"Artificial Intelligence Education Ethical Problems and Solutions",* que basadas en la Educación 4.0 sacan este artículo que traduce: Inteligencia artificial. Educación ética problemas y soluciones.

Este artículo tiene la finalidad de mostrar los problemas que surgen en la utilización de IA, ya que por tratarse de personas los datos no son siempre exactos, estos problemas se pueden dividir en 3 categorías: La irracionalidad del algoritmo; la incompletitud de los datos y la inexactitud del contenido, sin embargo quieren dar las posibles soluciones a esas variables; sin embargo descubrieron que la raíz de los problemas se debían a las personas, así que el estudio dividió a las personas y las categorizó en 3, según los diferentes aspectos de los que son responsables en la educación de inteligencia artificial como Administradores, profesores y estudiantes.

Este artículo resalta la importancia de la inteligencia artificial en la educación y como esta va evolucionando. La educación en inteligencia artificial es una nueva forma de educación. Su objetivo es utilizar la tecnología de inteligencia artificial para promover la reforma de los métodos de educación, mejorar el modelo de capacitación del personal y utilizar la tecnología de inteligencia artificial para mejorar el entorno de aprendizaje. En el Informe Horizon 2017 (Edición Básica), dice que la educación en inteligencia artificial tiene una influencia positiva

tanto en los maestros como en los estudiantes. La educación en inteligencia artificial es propicia para mejorar el nivel cognitivo de los estudiantes y reducir la carga de trabajo de los maestros.

Sin embargo, hay que rescatar que la inteligencia artificial está aún en una etapa inicial, y hay áreas que deben mejorarse, Entre esas áreas, los problemas más valiosos para resolver son los problemas éticos que surgen en la educación de la inteligencia artificial. Algunos problemas éticos traerán efectos negativos a la educación. Por ejemplo, las máquinas aceptarán todo tipo de información, incluso si hay algo mal, porque no tienen la capacidad de juzgar. Esto puede llevar a los robots a enviar información incorrecta a los estudiantes, o incluso darles valores incorrectos. A veces, los sistemas de inteligencia artificial también generan algunas asociaciones erróneas que conducen a decisiones irrazonables.

Además, es importante resaltar los conceptos de la inteligencia artificial y estudios en deferentes niveles de la educación como lo expresan en el artículo *"Artificial intelligence and computer science in education: From kindergarten to university",* existen módulos para enfocar e iniciar en temas fundamentales de IA como gráficos y estructuras de datos, clasificación y resolución de problemas por búsqueda, mientras que hay módulos más avanzados que cubren temas de IA como autómatas, agentes inteligentes, planificación y aprendizaje automático [5].

Así mismo otros artículos apoyados en la inteligencia artificial apoyan el trabajo de esta tesis con otros enfoques, incluyendo datos socio-demográficas y académicas para mejorar el rendimiento académico como género, estatus económico, tipo de escuela y desempeño [5] [6] . Además de lo anterior, otros indican que algunos cursos específicos sirven como indicadores significativos del rendimiento académico de los estudiantes y afirmaron que los cursos no son igualmente informativos para hacer predicciones precisas [26].

Con relación a las características de los individuos y sus hogares, el Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES) ha identificado variables que son cuantificables y permiten ver las diferencias en la tasa de deserción de acuerdo a las características de entrada de los jóvenes a la educación superior, de esta manera identificar las probabilidades de deserción de un estudiante matriculado en un programa de educación superior como lo son [27]:

- Estrato
- Sexo
- Nivel educativo de los padres.
- Ingresos económicos de la familia del estudiante.
- Nivel de clasificación del núcleo familiar según el SISBÉN .
- Número de Personas que componen el núcleo familiar.
- El joven trabajaba al momento de presentar las pruebas de Estado.
- Clasificación según los resultados de las pruebas de estado Saber 11°.
- Edad de Presentación de las Pruebas Saber 11°.

Estas variables no se tendrán en cuenta por lo que solamente se utilizaran variables académicas como nota de la materia, el semestre actual y cohorte a la que pertenece.

Este artículo [24], se refiere al estudio por medio de cadenas de Markov de unos estudiantes de la universidad Eafit, probando si era válido un modelo Markoviano enfocado a la evolución de cohortes de estudiantes, pero aplicado en particular a la forma como evolucionan los estudiantes a través de las materias y a través de una línea, un conjunto de materias que componen un área específica dentro de la formación profesional.

El estudio que se presenta utiliza el concepto de las cadenas de Markov. Una cadena de Markov como se ha mencionado anteriormente es el estudio de la sucesión de transiciones en el tiempo de un fenómeno aleatorio o estocástico, mediante el cual es posible estimar vectores de probabilidad de los estados del sistema en el futuro y el tiempo promedio de permanencia en cada estado. Los estados del sistema son los eventos en los que se encuentra el fenómeno, en cualquier instante de tiempo. El modelo, en esencia, estudia el flujo de un estudiante en una línea de materias, donde él en la primera materia de una línea puede: ganarla y avanzar a la siguiente materia; ganarla y salir temporalmente del programa (pedir reingreso o reintegro); ganarla y salir definitivamente del programa (deserción); perderla y repetirla; perderla y salir temporalmente del programa; perderla y salir definitivamente del programa.

En este estudio determinaron entonces, las proporciones de estudiantes que avanzan en cada materia, por cohorte, estas proporciones son estimaciones de las probabilidades de avanzar, de acuerdo con un enfoque frecuencial.

Sí bien la estimación de la demanda es un factor importante para la planeación de recursos, el conocimiento del tiempo promedio a la cual se desplazan los estudiantes en un programa académico, tiene implicaciones pedagógicas poco exploradas en nuestro medio y que son importantes para inferir en forma parcial sobre los factores que afectan dicha velocidad. La tendencia descendente del mayor rezago promedio por nivel, indica que las materias de ciencias básicas son aquellas que le presentan al estudiante mayor problema en su aprendizaje, ya sea por la característica propia de ellas, o porque las bases matemáticas con las que llegan de su bachillerato, presentan falencias.

También es cierto que en todo sistema, se requiere un tiempo de adecuación para llegar a la homogeneidad. La formación con la que los estudiantes llegan a la Universidad no es uniforme. Por lo tanto el rendimiento de los estudiantes en los primeros semestres reflejará esas diferencias en la calidad de la formación del bachillerato. Es entonces de esperar que haya una mayor fricción en ciertas materias de los semestres iniciales. En la práctica, si no existe rezago en estas materias, querría decir que los estudiantes o están pasando sin obstáculos (poca exigencia académica) o se tiene una excelente metodología pedagógica. En cualquiera de estas condiciones, es necesario establecer el por qué, sea para mejorar, o sea para mantener la excelencia de tales condiciones.

Otro de los antecedentes [28], en los que se apoya el presente trabajo de grado fue : El diseño de un modelo estocástico usando cadenas de Marcov para pronosticar la deserción académica de estudiantes de ingeniería en la Escuela de ingenieros Julio Garavito en el año 2012, utilizaron las cohortes del 2003 al 2006. Debido a la distribución de los datos, a que varias variables independientes son categóricas y a que la variable dependiente es una variable categórica ordenada, se empleó el modelo de regresión logística multinomial para obtener las probabilidades de transición, las que a su vez se usaron para modelar las cadenas de Markov absorbentes. Estas se utilizaron como una herramienta predictiva para determinar el tiempo de permanencia en la universidad, la probabilidad de graduarse, la probabilidad de deserción, e identificar las variables críticas que causan la deserción.

Con base en el planteamiento presentado, se trata de responder las siguientes preguntas: ¿Cuáles son las variables que afectan la deserción académica en el programa de Ingeniería de la Escuela Colombiana de Ingeniería Julio Garavito? ¿Cuál es el tiempo que permanecen los estudiantes

hasta graduarse? y ¿cuál es la probabilidad de deserción de estos, dadas unas características en su perfil como estudiante?.

Este es un estudio de caso cuantitativo y el diseño de la investigación es no experimental, longitudinal de evolución de grupo o cohorte, porque se toman los datos de cohortes completas y se analiza su comportamiento a lo largo del tiempo. No obstante, se hacen unas simulaciones con unos perfiles de estudiantes con características especiales para ver su comportamiento con el modelo estocástico que se diseñó, con el que se predicen el tiempo de permanencia y las probabilidades de graduarse y de retirarse de cualquiera de los programas de Ingeniería en la Escuela Colombiana de Ingeniería Julio Garavito.

Los estados de las cadenas de Markov se clasifican en:

Estado alcanzable. Dados dos estados i y j, una trayectoria de i a j es una secuencia de transiciones que comienza en i y termina en j, tal que cada transición en la secuencia tiene una probabilidad positiva de ocurrir. Un estado j es alcanzable desde el estado i si hay una trayectoria que conduzca de i a j.

Estados que se comunican. Se dice que dos estados i y j se comunican si j es alcanzable desde i, e i es alcanzable desde j.

Conjunto cerrado. Un conjunto de estados S de una cadena de Markov es cerrado si ningún estado fuera de S es alcanzable desde algún estado en S.

Estado absorbente. Es un estado en el que $p_{ii}=1$ y $p_{ij}=0$. Siempre que se entra en un estado absorbente, no se sale de él; un estado absorbente es un conjunto cerrado que contiene solo un estado.

Estado transitorio. Un estado i es transitorio si existe un estado j que es alcanzable desde i, pero el estado i no es alcanzable desde el estado j.

Estado recurrente. Es un estado que no es transitorio.

Estado periódico. Un estado i es periódico con periodo $k>1$ si k es el número más pequeño tal que las trayectorias que conducen al estado i de regreso al estado i tienen una longitud que es un múltiplo de k. Si un estado recurrente no es periódico, se conoce como periódico.

Cadenas absorbentes. Una cadena de Markov absorbente es una cadena en la cual algunos de sus estados son absorbentes y el resto son transitorios. Estas cadenas absorbentes tienen unas propiedades importantes que se explican a continuación.

La metodología que se usó en la investigación se sustenta en lo propuesto en el Modelo de flujos de educación estudiantil, usando cadenas de Markov. Es así como un estudiante va fluyendo a través de los semestres uno a uno, hasta alcanzar el grado. Una vez que ingresa a primer semestre, tiene tres posibilidades: la primera es que luego de cursar el semestre se retire de la institución, lo que para efectos de código interno se denomina SsR y se considera como una deserción; en el modelo de Markov es un estado absorbente, porque luego de que llega a este no se puede devolver.

Una segunda posibilidad es que el estudiante pierda el semestre, código interno SsP, y decida repetirlo; entonces sigue en el programa. Por último, existe la posibilidad de que el estudiante apruebe su semestre, código SsA, y avance sucesivamente hasta que se gradúe; el graduarse también es un estado absorbente.

Después del análisis realizado, es posible concluir que las variables que más afectan la deserción en la Escuela son el género del estudiante que sea hombre o mujer, las ciencias básicas, el estrato socio-económico, con quién vive actualmente y el tipo de vivienda.

Como se puede observar, en la Escuela Colombiana de Ingeniería no hay mucha población de los estratos 5 y 6 ni tampoco del 1, lo que predomina son los estratos 2, 3 y 4; así las cosas, los esfuerzos se deben orientar a lograr bajar la deserción en estos grupos de estudiantes.

Una investigación interesante, que podría ser la continuación del presente estudio, consistiría en averiguar cuáles de las asignaturas son las realmente críticas y cuál es la metodología que se usa, por ejemplo clases magistrales o no, exámenes conjuntos o no.

Llama la atención que estudiantes del estrato socio-económico 1 tengan una probabilidad más alta de graduarse. Valdría la pena hacer una investigación en la que se analizara qué motiva a estas personas a persistir a través del tiempo hasta lograr graduarse, así como también qué perfiles son los que perseveran y terminan sus estudios con éxito.

Otro antecedente es llamado: Simulador estocástico de rendimiento académico estudiantil basado en el método Monte Carlo [29], Este artículo realiza una predicción por medio de la probabilidad utilizando un histórico de datos a través del método Monte Carlo basado en métodos de programación estocástica que indica si un estudiante pasa o pierde, este modelo nos ayuda a desarrollar esta tesis de grado, ya que uno de los objetivos es poder predecir qué materias debe ver un estudiante para optimizar el tiempo marginal para que se gradué en el tiempo que es, para cesto se debe poder predecir qué materias se deben ver primero que otras para poder optimizar ese tiempo.

El propósito de este trabajo es el desarrollo de un simulador estocástico que determina el rendimiento de los estudiantes a lo largo de un plan de estudios dado, en este caso, Licenciatura en Tecnologías de Ingeniería, en la Universidad Técnica de Madrid.

El software desarrollado es genérico y puede adaptarse directamente a cualquier grado académico. Además, el usuario puede modificar el número y/o la definición de los indicadores de rendimiento si es necesario.

El modelo estocástico está basado en los siguientes supuestos:

1) Hay independencia entre el rendimiento académico de diferentes estudiantes, es decir, la probabilidad p que un estudiante s apruebe en la materia i no esta influenciada si el estudiante s' aprueba o no.

2) Para un estudiante dado, el rendimiento académico futuro es independiente de su rendimiento anterior, es decir, el número de intentos antes de pasar una materia i ($\left(k_i\right)$) no es influenciado directamente por el número de intentos de cualquier otra materia i' ($corr\left(k_i, k_{i'}\right)=0, \forall i \neq i'$)

3) La probabilidad de pasar la materia i en el intento es menor que la probabilidad de pasar la misma materia en el intento (En caso de fallar).

4) Se supone que los estudiantes se registraran en el semestre para: (i) las materias que fallaron en semestre anterior , y (ii) un conjunto de nuevas materias, sin sobrepasar el número de créditos permitido.

5) La probabilidad de pasar una materia (p_i) es modelada como una variable aleatoria uniformemente distribuida definida en el rango [p_i^{low} , p_i^{up}]. Este rango es determinado en el histórico de datos.

Corresponde a la probabilidad de pasar para la $i-th$ materia basada en el histórico de datos Se describe que el método Monte Carlo es utilizado cuando es engorroso o incluso inviable calcular los resultados exactos con un algoritmo determinístico.

El método de Monte Carlo empleado es:

1) Instancia de contadores: Contador de estudiantes $s \leftarrow 1$, Contador de año $y \leftarrow 1$, contador de semestre $t \leftarrow 1$. Instancia de conjuntos: $PS_s = FS_s = \varnothing$ como un conjunto de materias ganadas y fallidas. Instancia de probabilidad de pasar: Para cada materia la probabilidad p_i es determinada aleatoriamente con un límite definido por [p_i^{low} , p_i^{up}]

2) Matriculación del estudiante *s-th.* El estudiante es matriculado en las materias fallidas del conjunto FS_s, luego el estudiante es matriculado en el siguiente plan de estudio. Actualizar el conjunto de estudiantes registrados $RS_{i,t}$

3) Evaluación, El estudiante es evaluado estocásticamente de acuerdo a la probabilidad de la correspondiente materia, considerando el siguiente esquema. Conjunto de materias ganadas y fallidas es actualizado (PS_s y FS_s respectivamente)

4) Actualizar el contador $s \leftarrow s+1$. Si $s \leq 450$ entonces vaya al *paso 2).* De otra manera, actualizar $s \leftarrow 1$ y continuar.

5) Calcular los conjuntos de estudiantes que se presentaron y los que ganaron para el semestre t-th (y respectivamente). Actualizar el contador del semestre: $t \leftarrow t+1$. Si $t \leq 2$,

entonces vaya al *paso 2).* De otra manera, actualizar $t \leftarrow 1$ y continuar. Actualizar el contador de año: $y \leftarrow y+1$ y continuar.

6) Si s y el conjunto de materias aprobadas para todos los alumnos cumple con la planificación académica, el algoritmo finaliza. De otra manera vaya al *paso 2).*

Por otra parte se encontró un antecedente enfocado a los factores predictores del rendimiento académico [30]. Este trabajo de investigación se centró en investigar los factores de predicción del rendimiento académico de los estudiantes en la Universidad Mae Fah Luang, una universidad de reciente creación en Tailandia. Se proponen tres posibles factores: horas de actividad, puntajes de inglés y número de estudiantes admitidos.

Los resultados señalan que el éxito de los estudiantes podría ser estimulado a través de actividades fuera del aula, y que el rendimiento de la clase de inglés del primer año tiene un nivel relativamente alto de significativo para predecir el rendimiento académico general de los estudiantes.

La mayoría de los trabajos de Educational Data Mining (EDM) [31], se realizan a nivel de programa o curso; algunos generan modelos personalizados, mientras que esta investigación [30], se centró en modelos de predicción sobre el rendimiento académico en el nivel universitario.

Se inspecciono la siguiente información:

1) Horas individuales de actividades participadas y promedio de calificaciones de la escuela secundaria (GPA) individual en el primer semestre del primer año.

2) Los puntajes de Inglés 1 del individuo y el promedio de GPA del individuo en el cuarto año.

3) Total de número de estudiantes admitidos en cada año y el promedio de GPA del primer semestre del año de estos estudiantes. Algunos de estos datos que se proporcionaron de manera nominal debieron ser transformados a numéricos.

Como cada año ha habido varios tipos de admisiones, que tienen diferentes requisitos y procedimientos, la calidad de los estudiantes admitidos por diferentes tipos puede variar. Para obtener resultados más precisos, este conjunto de datos se dividió en 6 subconjuntos, cada uno de los cuales contiene información de los estudiantes admitida a través de enfoques similares.

Para estudiar las relaciones de los elementos propuestos y el rendimiento académico de los estudiantes, cada conjunto de datos se representó como un diagrama de dispersión.

Se encontró que todas las gráficas tenían la tendencia de relaciones lineales; por lo tanto, los modelos de regresión lineal podrían aplicarse en estos conjuntos de datos para fines de análisis. La regresión lineal simple explica la correlación de dos variables como una ecuación junto con algunos parámetros. En otras palabras, predecirá una variable llamada variable dependiente (Y) , sobre la información de la otra variable, llamada variable independiente (x) como se muestra a continuación:

$$\hat{Y} = b_0 + b_1 x \qquad \textbf{(2.2.1)}$$

Donde b_0 y b_1 son variables no conocidas que pueden ser adquiridas desde un análisis de regresión.

El rendimiento académico de los estudiantes se establece como y los factores propuestos se establecen como en cada modelo, donde cada par de datos contribuyó a un modelo excepto el par de datos (3) que generó un modelo a partir de cada conjunto de datos por subgrupos explicados anteriormente. Además, se eligió el nivel significativo a 0.05.

H_{01}: La cantidad de horas de actividad que un estudiante pasa no Tiene un impacto en el rendimiento académico del alumno. El valor p-value es 0, que es menor que el nivel significativo elegido, por lo tanto, la hipótesis se rechaza. Esto significa que el total de horas de actividad es estadísticamente significativo y representa el 5.46% del rendimiento académico del estudiante en el primer semestre.

H_{02}: El grado de Inglés 1 no determina el GPA general de un estudiante. Se revela que el rendimiento en la clase de Inglés 1 es un factor predictivo importante del GPA general de un estudiante al final del programa de estudio con una influencia del 20.64% y la correlación de 0.454.

El conocimiento confirma que las buenas habilidades de inglés son necesarias para estudiar en MFU o en cualquier institución donde el inglés se usa como idioma principal para la enseñanza y el aprendizaje con el fin de obtener las calificaciones deseadas en los cursos.

H_{03}: El número total de estudiantes admitidos no afecta el rendimiento académico de los estudiantes. Para los tipos de admisión 1 y 2, con el valor p de 0.539 y 0.407 respectivamente, no existe una correlación estadísticamente significativa entre el número de estudiantes admitidos y el rendimiento académico de los estudiantes.

Para los tipos de admisión 3 y 6, se mostró que el número de estudiantes tiene una débil correlación de 0.172 y 0.209 y representa el 2.96% y el 4.44% del rendimiento académico,

respectivamente. Así mismo dicen que la relación lineal positiva entre el número de estudiantes aceptados en la cuota de la provincia del Norte y el promedio de calificaciones (GPA) correspondiente. Sin embargo, la correlación de la explicación de 0.206 y 4.3% es débil.

La relación positiva podría aclararse por el hecho de que los estudiantes y sus padres confían más en la universidad, ya que se ha ganado más reputación a lo largo de los años, lo que hace que más estudiantes con talento en las provincias cercanas se registren en MFU en lugar de irse más lejos de casa.

Los resultados que obtuvieron demuestran que la correlación relativamente fuerte es de 0.392 y el coeficiente de determinación correspondiente es del 15,3%, reflejan las características de este tipo de admisión.

En el antecedente: "Factores predictores del rendimiento académico" [26], se crean diferentes modelos de clasificación para predecir el rendimiento de los estudiantes, utilizando datos recogidos de una universidad Australiana.

La colección de datos utilizados incluye el detalle de matrícula del estudiante, así como los datos de la actividad generada desde el sistema de administración de aprendizaje de la universidad. Los datos de matrícula contienen información del estudiante, como características socio-demográficas, base de admisión a la universidad (por ejemplo, a través de un examen de ingreso o experiencia pasada) y tipo de asistencia (por ejemplo, tiempo completo vs. tiempo parcial). Y los datos obtenidos del Learning management system (LMS) - Moodle registran la participación del estudiante en diferentes actividades (por ejemplo, tareas, exámenes, foros y otros) y recursos (por ejemplo, libros y archivos).

Una contribución importante de este estudio es la consideración de la heterogeneidad de los estudiantes en la construcción de modelos predictivos. Su principal objetivo es analizar la diferencia existente en las características de un estudiante, considerando características demográficas y académicas claves para influenciar en el rendimiento académico. Por lo tanto, el estudio tiene como objetivo construir modelos de predicción en diferentes subpoblaciones, teniendo en cuenta el género, la edad y el tipo de asistencia del estudiante.

Para tener un enfoque en dos términos como la capacidad de predicción y la interpretabilidad del desarrollo de los modelos, se aplican dos métodos de caja negra y dos de caja blanca para generar submodelos. Los métodos caja negra son naive-Bayes y sequential minimal optimizer (SMO), mientras que los de caja blanca son J48 y JRip. La capacidad para predecir el rendimiento del estudiante e identificar el riesgo de los estudiantes de fracasar es un área de investigación que está en crecimiento. Técnicas de Minería de datos han sido aplicadas con éxito para predecir el rendimiento académico del estudiante.

El preprocesamiento de datos es una fase importante para preparar los datos antes de aplicar los métodos de extracción de datos. (i) Todos los atributos de actividad se clasifican en cuatro cuartiles, a saber, Q1, Q2, Q3 y Q4, donde Q1 representa la participación más baja y Q4, la más alta. (ii) Se realizó la transformación del formato de los datos de Excel a ARFF para aplicar los métodos de clasificación con un formato permitido.

La partición del conjunto de datos se realizó en dos pasos, (i) la inscripción, la actividad y los conjuntos de datos combinados se dividen de acuerdo con el género del estudiante (masculino y femenino), la edad (normal y madura), el tipo de asistencia (a tiempo completo y parcial) y el modo de asistencia (interno y externo) y se generan 8 subconjuntos de datos para cada uno de los

conjuntos de datos de inscripción, actividad y combinados, respectivamente. (ii) los subconjuntos de datos femeninos y masculinos se dividen en otros 6 conjuntos de datos según la edad del alumno, el tipo de asistencia y el modo de asistencia. Para todas las ejecuciones se usó WEKA. Para la predicción del rendimiento académico se emplearon cuatro métodos de clasificación para generar submodelos de estudiantes. (i) Naïve-Bayes: Este es un método de clasificación probabilística basado en el teorema de Bayes, este clasificador puede usarse con atributos discretos o continuos. (ii) SMO: Este método utiliza un algoritmo de optimización para entrenar una máquina de vectores de soporte (SVM). (iii) J48: Este método genera un árbol de decisión que contiene tres tipos de nodos diferentes: raíz, interno y nodos de hoja; el nodo raíz es el nodo más alto de un árbol. La raíz y los nodos internos contienen condiciones de prueba de atributos, donde cada rama representa un resultado de la prueba y cada nodo de hoja representa un nivel de clase. Los clasificadores basados en árboles de decisión exhiben alta precisión y son fáciles de implementar. (iv) JRip: Este es un método de clasificación basado en reglas que genera reglas comprensibles en una estructura IF-THEN. En la evaluación de los modelos generados han sido desarrollados un número de criterios para medir la capacidad de predicción del modelo. (i) Precisión: la fracción de ejemplos positivos verdaderos entre todos los ejemplos clasificados como positivos por un clasificador. (ii) Recall: La fracción de ejemplos positivos verdaderos clasificados correctamente por un clasificador. (iii) F-measure: la media armónica de la precisión y recall de un clasificador; es decir, medida F = 2 × precisión × recall / (precisión + recall). (iv) Kappa co-efficient: compara la precisión de un clasificador con la precisión que se espera que logre un clasificador aleatorio. (v) AUC: el área bajo la curva del receiver operating

characteristic (ROC) indica la probabilidad de que un clasificador clasifique un ejemplo positivo seleccionado al azar más que un ejemplo negativo seleccionado aleatoriamente.

Para ir finalizando los antecedentes para el presente trabajo de grado se tuvo en cuenta "Aplicación de métodos de análisis de aprendizaje para mejorar la calidad y efectividad del aprendizaje en entornos virtuales de aprendizaje" [32]. El objetivo del artículo es analizar la aplicación de los métodos de análisis de aprendizaje (LA) y Educational Data Mining (EDM) para mejorar la calidad y la eficacia del aprendizaje en entornos virtuales de aprendizaje (VLE) a través de la personalización del aprendizaje.

Una de las tendencias claras es el aumento constante del número de artículos publicados de 2008 a 2017. LA es el análisis de los datos de aprendizaje electrónico que permite a los maestros, diseñadores de cursos y administradores de VLE buscar patrones no observados e información subyacente en los procesos de aprendizaje. El objetivo principal de LA es mejorar los resultados de aprendizaje y el proceso de aprendizaje general en las aulas virtuales de aprendizaje electrónico y la educación asistida por computadora.

El rol de la DM se ha vuelto crucial para extraer esta información oculta de big data. Las técnicas básicas en LA / EDM y sus ejemplos de aplicación son:

1) Clasificación (para clasificar cada elemento en un conjunto de datos en uno de un conjunto predefinido de grupo de alumnos),

2) Clustering (para determinar grupos de estudiantes que necesitan perfiles de cursos especiales);

3) Association rules (para descubrir relaciones interesantes entre los elementos del curso que fueron utilizados por el estudiante),

4) Predicción (para predecir dependencias en el uso del contenido y las actividades del sistema de aprendizaje), y

5) Árboles de decisión (realizan muchas funciones de análisis de aprendizaje y se pueden usar en una variedad de tareas, como la previsión o el análisis).

Estas técnicas se pueden utilizar juntas o una después de otra, según la complejidad de la tarea resuelta.

En primer lugar, los objetos de aprendizaje basados en VLE y los métodos / actividades de aprendizaje deben estar interconectados con los estilos de aprendizaje de los estudiantes, por ejemplo, El modelo de estilos de aprendizaje de Honey y Mumford utiliza métodos de evaluación expertos basados en, por ejemplo, Números difusos (fuzzy numbers) o proceso de jerarquía analítica.

En la segunda etapa, se analiza un grupo de estudiantes para identificar sus estilos de aprendizaje individuales.

En la tercera etapa, utilizando los métodos apropiados de LA / EDM identificados en la sección anterior, podríamos analizar qué objeto de aprendizaje particular o métodos / actividad de aprendizaje fueron utilizados prácticamente por estos estudiantes en VLE, y en qué medida.

Los datos sobre el uso práctico de los objetos de aprendizaje basados en VLE y los métodos / actividades deben compararse con los índices probabilísticos de idoneidad de los estudiantes (etapa IV). Las rutas de aprendizaje personales de los estudiantes en VLE se deben corregir de acuerdo con los nuevos datos identificados (etapa V).

Etapas I – II: el modelo de estilos de aprendizaje de Honey y Mumford identificó cuatro estilos o preferencias de aprendizaje diferentes: Activista, Teórico; Pragmático y Reflector: (1) Activista

(ACT): los activistas son aquellas personas que aprenden haciendo. (2) Reflector (REF): estas personas aprenden observando y pensando en lo que sucedió. (3) Pragmático (PRG): estas personas necesitan poder ver cómo poner en práctica el aprendizaje en el mundo real. (4) Teórico (THR): A estos estudiantes les gusta entender la teoría detrás de las acciones.

En función de los resultados de completar el correspondiente cuestionario psicológico dedicado, el perfil de un alumno se podría modelar como un conjunto de criterios múltiples $L = LSt(\{w1, w2, w3, w4\})$, donde $\{w1, w2, w3, w4\}$ son los valores del estilo de aprendizaje de acuerdo con la tipología elegida de los estilos de aprendizaje, donde $w1: Activista, w2: Refector, w3: Pragmatico, w4: Teorico\}$.

Los métodos y técnicas de Learning Analytics se podrían utilizar con éxito para personalizar el aprendizaje en entornos de aprendizaje virtuales. La clasificación, agrupación, reglas de asociación, predicción y árboles de decisión son los métodos / técnicas de Learning Analytics más adecuados para aplicar a la personalización del aprendizaje.

Finalmente tenemos un antecedente final. Predecir a los estudiantes en buen estado al correlacionar los rasgos de personalidad relevantes con los datos académicos y profesionales [33]. Este documento presenta el progreso de los resultados en los algoritmos Good Fit Students (GFS) y la construcción matemática. Este trabajo aborda los problemas de rendimiento académico deficiente, bajos índices de retención, abandonos escolares, transferencias escolares, reingresos costosos, rendimiento deficiente, transferencias tempranas de trabajo y utilización / consideración ineficiente de talentos naturales.

Los algoritmos fueron desarrollados para predecir la relevancia educacional de talentos individuales a través de características personales (Estructurados y desestructurados) y los datos de la carrera académica.

3 CAPÍTULO III: MARCO CONCEPTUAL

3.1. Fundamentación teórica

3.1.1 Cadenas de Markov.

Las cadenas de Markov son un modelo de probabilidad adecuado para ciertas series de tiempo en las que la observación en un momento dado es la categoría en la que cae un individuo [3].

La cadena de Markov más simple es aquella en la que hay un número infinito de estados o categorías y un número infinito de puntos de tiempo equidistantes en los que se realizan observaciones, las cadenas son de primer orden y las probabilidades de transición son las mismas para cada intervalo de tiempo. Dicha cadena se describe por el estado inicial y el conjunto de probabilidades de transición, la probabilidad condicional para entrar en cada estado, dado el estado inmediatamente anterior.

También hay una variable, y es la que Bartlett y Hoel han estudiado [28], el problema de la estimación de las probabilidades de transición y de la prueba de la bondad del ajuste y el orden de la cadena en la situación en la que solo se observa una secuencia de estados.

En esta investigación [3], se encuentra la estimación de los parámetros de una cadena de Markov de primer orden, donde el modelo deja que los estados sean: $i=1,2,\ldots,m$. Aunque el estado i es generalmente considerado como un número entero de 1 a m , no se hace uso real de

esta disposición ordenada, por lo que podría ser, por ejemplo, un partido político, un lugar geográfico, un par de número (a,b) , entre otros.

Los tiempos de observación son $i=0,1,...,T$ y $p_{ij}(i,j=1,\ldots,m;t=1,\ldots,T)$ es la probabilidad del estado j en el tiempo t , dado el estado i en el tiempo $t-1$. Trataremos tanto con (a) probabilidades de transición estacionarias (es decir, $p_{ij}(t)=p_{ij}$ para $t=1,...,T$) como con (b) probabilidades de transición no estacionarias (es decir, donde las probabilidades de transición no necesitan ser las mismas para cada intervalo de tiempo). Se supone que hay $n_i(0)$ individuos en el estado i para $t=0$. Se supone además que los $n_i(0)$ como si fueran no aleatorios. Una observación sobre un individuo, consiste en la secuencia de estados en los que se encuentra el individuo $i=0,1,...,T$, es decir, $i(0),i(1),i(2),\ldots,i(T)$. Dado el estado inicial $i(0)$, hay m^T secuencias posibles. Estos representan eventos mutuamente excluyentes con probabilidades [3].

$$p_{i(0)i(1)}\,p_{i(1)i(2)}\cdots p_{i(T-1)i(T)} \qquad \textbf{(3.1.1.1)}$$

Cuando las probabilidades de transición son estacionarias. (Cuando las probabilidades de transición no son necesariamente estacionarias, los símbolos de la forma $p_{i(t-1)i(t)}$ deben reemplazarse por $p_{i(t-1)i(t)}(t)$ [34].

Así $n_{ij}(t)$ denota el número de individuos en el estado i para $t-1$ y j para t . Se demuestra que el conjunto de $n_{ij}(t)(i,j=1,\ldots,m;t=1,\ldots,T)$ es un conjunto de m^{2T} números, forma un conjunto de estadísticas suficientes para las secuencias observadas. Sea

$n_{i(0)i(1)\ldots i(T)}$ el número de individuos cuya secuencia de estados es $i(0),\ldots,i(T)$ [3]. Entonces:

$$n_{gj}(t)=\sum n_{i(0)i(1),\ldots,i(T)} \qquad \textbf{(3.1.1.2)}$$

Donde la suma está sobre todos los valores de $i's$ con $i(t-1)=g$ y $i(t)=j$. La probabilidad, en el espacio dimensional nmT que describe todas las secuencias para todos los n individuos (para cada estado inicial hay nT dimensiones) de un conjunto o secuencias ordenadas para n individuos:

$$\begin{aligned}
&\prod [p_{i(0)i(1)}(1)p_{i(1)i(2)}(2)\ldots p_{i(T-1)i(T)}(T)]^{n_{i(0)i(1)\ldots i(T)}} \\
&=(\prod [p_{i(0)i(1)}(1)]^{n_{i(0)i(1)\ldots i(T)}})\ldots(\prod [p_{i(T-1)i(T)}(T)]^{n_{i(0)i(1)\ldots i(T)}}) \\
&=(\prod_{i(0),i(1)} p_{i(0)i(1)}(1)^{n_{i(0)i(1)}(1)})\ldots(\prod_{i(T-1),i(T)} p_{i(T-1)i(T)}(T)^{n_{i(T-1)i(T)}(T)}) \qquad \textbf{(3.1.1.3)} \\
&=\prod_{t=1}^{T}\prod_{g,j} p_{gj}(t)^{n_{gj}(t)}
\end{aligned}$$

Donde los productos en las dos primeras líneas están sobre todos los valores de los índices $T+1$. Así, el conjunto de números $n_{ij}(t)$ forma un conjunto de estadísticas suficientes. La distribución real de $n_{ij}(t)$ se multiplica por (5.2.1.3) por una función apropiada de factoriales

[3]. Así $n_i(t-1)=\sum_{j=1}^{m} n_{ij}(t)$, la distribución condicional de $n_{ij}(t), j=1,\ldots,m$, dado $n_i(t-1)$ (o dado $n_k(s), k=1,\ldots,m; s=0,\ldots,t-1$) es:

$$\frac{n_i(t-1)!}{\prod_{j=1}^{m} n_{ij}(t)!}\prod_{j=1}^{m} p_{ij}(t)^{n_{ij}(t)} \qquad \textbf{(3.1.1.4)}$$

Estas son las mismas distribuciones que se obtendrían si se tuvieran $n_i(t-1)$ observaciones en una distribución multinomial con probabilidades $p_{ij}(t)$ y con los números resultantes $n_{ij}(t)$. Las distribuciones de $n_{ij}(t)$ (condicional en $n_i(0)$) es:

$$\prod_{t=1}^{m}\left\{\prod_{i=1}^{m}\left[\frac{n_i(t-1)!}{\prod_{j=1}^{m} n_{ij}(t)!}\prod_{j=1}^{m} p_{ij}(t)^{n_{ij}(t)}\right]\right\} \qquad \textbf{(3.1.1.5)}$$

Cuando la probabilidad de transición es estacionario, la probabilidad puede ser escrita de la siguiente forma [3]:

$$\prod_{t=1}^{T}\prod_{g,i} p_{g,j}^{n_{g,i}(t)}=\prod_{i,j} p_{i,j}^{n_{i,j}} \qquad \textbf{(3.1.1.6)}$$

3.1.2 Máximo estimado de verosimilitud.

Las probabilidades de transición estacionarias p_{ij} pueden ser estimada maximizando la probabilidad (5.2.1.6) con respecto a la p_{ij}, sujeto a las restricciones $p_{ij}\geq 0$ y

$$\sum_{j=1}^{m} p_{ij}=1, \quad i=1,2,\ldots,m. \qquad \textbf{(3.1.2.7)}$$

Cuando los n_{ij} son las observaciones reales. Esta probabilidad es exactamente de la misma forma, excepto por un factor que no depende de p_{ij} como el obtenido para m muestras independientes, donde la $i-ésima$ muestra $(i=1,2,...,m)$ consiste en $n_i^* = \sum_j n_{ij}$ ensayos multinomiales con probabilidades $p_{ij}(i,j=1,2,\ldots,m)$. Para tales muestras, es bien conocido y se verifica fácilmente que las estimaciones de máxima verosimilitud para p_{ij} son [3]:

$$\widehat{p_{ij}} = n_{ij}/n_i^* = \sum_{t=1}^{T} n_{ij}(t) / \sum_{k=1}^{m} \sum_{j}^{T} n_{ik}(t) = \sum_{t=1}^{T} n_{ij}(t) / \sum_{t=0}^{T-1} n_i(t) \qquad \textbf{(3.1.2.8)}$$

y, por lo tanto, esto también es cierto para cualquier otra distribución en la que la probabilidad elemental sea de la misma forma, excepto para los factores libres de parámetros, y las restricciones en p_{ij} son las mismas. En particular, se aplica a la estimación de los parámetros p_{ij} en (5.2.1.6).

Cuando las probabilidades de transición no son necesariamente estacionarias, el enfoque general utilizado en el párrafo anterior todavía se puede aplicar, y se encuentra que las estimaciones de probabilidad máxima para el $p_{ij}(t)$ [3]son:

$$\widehat{p_{ij}}(t) = n_{ij}(t)/n_i(t-1) = n_{ij}(t) / \sum_{k=1}^{m} n_{ik}(t) \qquad \textbf{(3.1.2.9)}$$

Las mismas estimaciones de máxima verosimilitud para la $p_{ij}(t)$ se obtienen cuando consideramos la distribución condicional de $n_{ij}(t)$ dado $n_i(t-1)$ como cuando se utiliza la

distribución conjunta de $n_{ij}(1), n_{ij}(2), \ldots, n_{ij}(T)$. Formalmente, estas estimaciones son las mismas que se obtendrían si para cada i y j se tuvieran observaciones $n_i(t-1)$ en una distribución multinominal con probabilidades $p_{ij}(t)$ y con los números resultantes $n_{ij}(t)$.

Las estimaciones se pueden describir de la siguiente manera: Las entradas $n_{ij}(t)$ para un determinado t se ingresen en una tabla bidireccional $m \times m$. La estimación de $p_{ij}(t)$ es la entrada $i, j-ésima$ en la tabla dividida por la suma de las entradas en la $i-ésima$ fila. Para estimar $p_{ij}(t)$ una cadena estacionaria, agregue las entradas correspondientes en las tablas de dos vías para $t = 1, \ldots, T$ obteniendo una tabla de dos vías con entradas $n_{ij} = \sum_t n_{ij}(t)$ La estimación de $p_{ij}(t)$ es la entrada $i, j-ésima$ de la tabla de n_{ij} dividida por la suma de las entradas en la $i-ésima$ fila [3].

3.1.3 Sistema de colaboración de filtrado.

Existen diferentes sistemas de recomendación con deferentes técnicas, este estudio se centrara en uno de los métodos tradicionales, como lo es el método basado en el filtrado colaborativo. Además, encontramos otros métodos como el filtrado por el contenido, el conocimiento y los híbridos además los métodos avanzados desarrollados recientemente, como enfoques basados en conjuntos difusos, basados en redes sociales, basados en la confianza, basados en el contexto y recomendaciones grupales [35].

Las técnicas de recomendación basadas en el filtrado colaborativo (CF) ayudan a las personas a tomar decisiones basadas en las opiniones de otras personas que comparten intereses similares

[36]. La técnica de CF se puede dividir en enfoques de CF basados en el usuario y basados en elementos. En el enfoque de CF basado en el usuario, un usuario recibirá recomendaciones de artículos que le gusten a usuarios similares. En el enfoque de CF basado en elementos, un usuario recibirá recomendaciones de elementos que son similares a los que ha tomado en el pasado.

La similitud entre usuarios o elementos se puede calcular mediante la similitud basada en la correlación de Pearson, la similitud basada en la correlación de Pearson restringida (CPC), la similitud basada en el coseno o las medidas ajustadas basadas en el coseno. Al calcular la similitud entre los elementos que utilizan las medidas anteriores, solo se consideran los usuarios que han calificado ambos elementos. Esto puede influir en la precisión de similitud cuando los artículos que han recibido un número muy pequeño de calificaciones expresan un alto nivel de similitud con otros artículos [37].

Hay un proceso que permite mejorar la precisión de similitud, utilizando un enfoque CF mejorado basado en elementos, combinando el enfoque coseno ajustado con la métrica Jaccard como un esquema de ponderación. Para calcular la similitud entre usuarios, la métrica Jaccard se utiliza como un esquema de ponderación con el CPC para obtener una medida de CPC ponderada. Para abordar la desventaja del enfoque basado en la calificación única, se desarrolla el filtrado colaborativo multicriterio [4].

4 CAPÍTULO IV: DESARROLLO DE LA TESIS

4.1. Nomenclatura

Tabla I. Variables del Modelo Formal

No.	Nombre	Descripción	Definición	Tipo	Ejemplo
(4.1)	ISC	Programa de ISC en jornada diurna	Nombre abreviado que permite identificar rápidamente al programa de ISC	Abrev.	Programa de ISC de la UTP
(4.2)	m	Identificador único de materia	Identifica una materia del programa de ISC en la UTP	Parámetro	$BA172$, $CB115$
(4.3)	st	Estudiante del programa de ISC de la UTP	Identificación simulada dada aleatoriamente por el sistema para identificar un estudiante	Parámetro	st_1
(4.4)	k	Número total de semestres	Cantidad de semestres en el que se matriculo un estudiante en la UTP	Parámetro	$1,\ldots,k$
(4.5)	$k+1$	Siguiente semestre	Siguiente semestre que tomara el estudiante	Parámetro	$1,\ldots,k,k+1$
(4.6)	m_k	Materia	Materia m matriculada en el semestre k	Parámetro	$(IS105)_3$

Nota: se asignan las variables necesarias para el desarrollo del modelo matemático, así mismo se adicionan los parámetros y las abreviaturas utilizadas para la definición de las variables.

Tabla I. (continuación)

No.	Nombre	Descripción	Definición	Tipo	Ejemplo
(4.7)	$m_{(k+1)}$	Materia	Materia $m_{(k+1)}$ matriculada después de tomar las materias consecutivas $\{m_1,\ldots,m_k\}$ en k semestres anteriores	Parámetro	$(CB314)_{(3+1)}$
(4.8)	s	Estado en el modelo de cadena de Markov	Conjunto de materias matriculadas en k semestres consecutivos $m_1,\ldots,m_k$	Parámetro	$(BA172)_1,\ldots$ $(IS105)_3$
(4.9)	s_1	Estado actual	Conjunto de materias matriculadas en k semestres consecutivos $m_1,\ldots,m_k$	Parámetro	$(BA172)_1,\ldots$ $(IS105)_3$
(4.10)	s_2	Estado siguiente	Conjunto de materias matriculadas en $k+1$ semestres consecutivos $m_1,\ldots,m_{(k+1)}$	Parámetro	$(BA172)_1,\ldots$ $(IS105)_3,$ $(CB314)_4$
(4.11)	$count_{st}$	Contador de estudiantes	Contador de estudiantes que coinciden con algún parámetro especifico.	Contador	$count_{st}$ $(Parámetro)$

Tabla I. (continuación)

No.	Nombre	Descripción	Definición	Tipo	Ejemplo
(4.12)	$countSt_{m_{k+1}}$	Contador	Indica el número de estudiantes que tomaron $m_{(k+1)}$ después de tomar las materias consecutivas $\{m_1, m_2, m_3, \ldots, m_k\}$ en k	Función	$count_{st}(s_1 \rightarrow (CB314)_4)$
(4.13)	$countSt_{m_k}$	Contador	Es el número total de estudiantes que tomaron $\{m_1, m_2, \ldots, m_k\}$ en k semestres consecutivos	Función	$count_{st}(s_1)$
(4.14)	MLE	Estimación de Máxima Verosimilitud $p(s_2 \mid s_1)$	Se refiere a la máxima probabilidad que se puede obtener por medio de la estimación de los datos de inscripción en materias	Formula	$\frac{countSt_{m_{k+1}}}{countSt_{m_k}}$
(4.15)	j	Semestre	El semestre j para el cual será calculado el puntaje de recomendación	Parámetro	4
(4.16)	m_j	Materias	Materias probables a matricular por un estudiante st en el semestre j	Parámetro	$(CB314)_4$

Tabla I. (continuación)

No.	Nombre	Descripción	Definición	Tipo	Ejemplo
(4.17)	s_a	Secuencias de materias tomadas por un estudiante	s_a es una secuencia de materias consecutivas tomadas por el estudiante en k (8.4) semestres anteriores. $\{m_{(j-1)},\ldots,m_{(j-k)}\}$	Parámetro	$(BA172)_1,\ldots$ $(IS105)_3$
(4.18)	$s_a \rightarrow s_a \cup\{m_j\}$	Unión de materias	Es la unión de s_a y las materias probables a matricular m_j	Formula	$s_a \rightarrow s_a \cup\{(CB314)_4\}$
(4.19)	$r(st, m_j, j)$	puntaje de recomendación	Puntaje de recomendación para cada materia m_j que un estudiante st es probable que tome en el semestre j	Formula	Suma de probabilidades de recomendación
(4.20)	n	Número de semestres	Semestres consecutivos en la construcción de una cadena de materias	Parámetro	2
(4.21)	λ	Peso	Es un coeficiente entre 0 y 1	Parámetro	1/2
(4.22)	$W(st, s_1)$	Peso	Es el peso asignado al estado $\{m_1,\ldots,m_k\}$	Formula	$W(st, s_1)$
(4.23)	$W(st, s_2)$	Peso	Es el peso asignado al estado $\{m_1,\ldots,m_k,m_{(k+1)}\}$	Formula	$W(st, s_2)$

4.2. Modelo matemático

Inicialmente se aplicara un modelo de Cadenas de Markov orientado al progreso de cohortes de estudiantes, pero utilizado en específico a la manera como progresan los estudiantes a través de grupos de materias por semestre que hacen parte de un dominio especifico en el proceso de formación profesional, los modelos implementados se desarrollaron basados en el modelo básico de Markov y por omisión [7].

4.2.1 Modelo básico de Markov

Para ordenar las materias en el modelo de filtrado de recomendación de materias, se modeló la secuencia de materias que un estudiante matricula, como una cadena de Markov en el cual las materias que un estudiante matriculará en el semestre $k+1$ (4.5) dependerá únicamente de las materias que aprobó en los k (4.4) semestres anteriores. El conjunto de k materias matriculadas en k semestres consecutivos $s=\{m_1,\ldots,m_k\}$ (4.8), identifica un estado en el modelo básico de cadenas de Markov.

La probabilidad de transición esta dada desde el estado $s_1=\{m_1,\ldots,m_k\}$ (4.9) al estado $s_2=\{m_1,\ldots,m_{(k+1)}\}$ (4.10) y puede ser estimada por medio de los datos de inscripción en materias utilizando la Estimación de Máxima Verosimilitud (4.14) [3] como se muestra a continuación:

$$p(s_2=\{m_1,\ldots,m_k,m_{(k+1)}\}|s_1=\{m_1,\ldots,m_k\})=\frac{count_{st}(\{m_1,\ldots,m_k\}\rightarrow m_{(k+1)})}{count_{st}(\{m_1,\ldots,m_k\})} \qquad \textbf{(4.2.1.24)}$$

Donde $count_{st}(\{m_1,\ldots,m_k\}\rightarrow m_{(k+1)})$ (4.12) indica el número de estudiantes que matricularon $m_{(k+1)}$ (4.7) después de matricular las materias consecutivas $\{m_1,\ldots,m_k\}$ (4.9) en k (8.4) semestres anteriores, y $count_{st}(\{m_1,\ldots,m_k\})$ (4.13) es el número total de estudiantes que matricularon $\{m_1,\ldots,m_k\}$ (4.9) en k (4.4) semestres consecutivos.

Normalmente los estudiantes se matriculan en mas de una materia por semestre, cada estudiante es asignado a varios estados en el espacio de estados correspondientes a varias combinaciones de materias que ha tomado en k (4.4) ó $(k+1)$ (4.5) semestre consecutivo.

Se calcula un puntaje de recomendación $r(st,m_j,j)$ (4.19) para cada materia m_j (4.16) que un estudiante st (4.3) es probable que tome en el semestre j (4.15) dadas sus inscripciones en k (4.4) semestres anteriores $s_a=m_{(j-1)},\ldots,m_{(j-k)}$ (4.17), como se muestra a continuación:

$$r(st,m_j,j)=\sum_{s_a=\{m_{(j-1)},\ldots,m_{(j-k)}\}} p(s_a\cup\{m_j\}|s_a) \qquad \textbf{(4.2.1.25)}$$

Esta fórmula anterior suma todas las probabilidades de transición para $s_a\rightarrow s_a\cup\{m_j\}$ (4.18) donde s_a (4.17) es una secuencia de materias consecutivas matriculadas por el estudiante en k (4.4) semestres anteriores.

Tabla II. Muestra de datos de inscripción de cuatro estudiantes en tres semestres consecutivos.

Estudiante	Semestre	Materias
st_1 : 1255820000280	1	BA172, CB115
	2	BA372, CB234
	3	BUB1, IS105
st_2 : 1259498737788	1	BA172, IS105
	2	BA372, CB234, CB242
	3	CB314, CB334
st_3 : 1267745842136	1	IS105, IS142
	2	BA172, CB115
	3	BA372, CB234
st_4 : 1278685878516	1	BA172, CB115
	2	BA372, IS284
	3	CB215, CB234

Nota: ilustra la implementación del modelo básico de Markov, se utiliza una muestra del conjunto de datos de matrículas de cuatro estudiantes en tres semestres consecutivos.

4.2.2 Modelo por omisión en cadenas de Markov

En el modelo matemático propuesto en el punto anterior es posible que algunos estudiantes no tengan recomendación de materias, esto nos lleva a investigar otros modelos matemáticos que han sido desarrollados específicamente para extraer información importante de un conjunto de datos con limitación de información en las cadenas de Markov que buscan materias en semestres consecutivos [7].

Este modelo por omisión permite aceptar que las materias que un estudiante debería matricular en semestres $(k+1)$ (4.5) no dependen únicamente de los k (4.4) semestres anteriores, pero también puede depender de los semestres anteriores. Por lo tanto, para mantener este modelo sin memoria de los semestres consecutivos, se permite omitir algunos semestres para la construcción del conjunto de materias $s_1=\{m_1,\ldots,m_k\}$ (4.9) y $s_2=\{m_1,\ldots,m_{(k+1)}\}$ (4.10).

Para diferenciar los estados que han sido construidos con y sin omisión se asigna un peso a cada estado. Entre más semestres sean omitidos en un estado menos peso tiene la predicción de la materia que debería matricular en el siguiente semestre. Cuando se omiten n (4.20) semestres consecutivos en la construcción de una cadena de materias, es necesario que a la cadena se le asigne un peso igual a λ^n , donde λ (4.21) es un coeficiente entre 0 y 1. Por ejemplo, se asigna un peso igual a λ si un semestre es omitido en la construcción de una cadena de materia, y se asigna un peso igual a λ^2 si dos semestres consecutivos son omitidos en la cadena. El peso de λ seria igual a 1 si no se omiten semestres consecutivos. La probabilidad de transición de este modelo se calcula con la siguiente formula:

$$P(s_1=\{m_1,\ldots,m_k\}\rightarrow s_2=\{m_1,\ldots,m_k,m_{(k+1)}\})=\frac{\sum_{st} W(st,\{m_1,\ldots,m_k,m_{(k+1)}\})}{\sum_{st} W(st,\{m_1,\ldots,m_k\})} \qquad \textbf{(4.2.1.26)}$$

Donde $W(st,\{m_1,\ldots,m_k\})$ (4.22) es el peso asignado al estado $\{m_1,\ldots,m_k\}$ (8.9) y $W(st,\{m_1,\ldots,m_k,m_{(k+1)}\})$ (4.23) es el peso asignado al estado $\{m_1,\ldots,m_k,m_{(k+1)}\}$ (4.10) para cada estudiante st (4.3).

El puntaje de recomendación $r(st, m_j, j)$ (4.19) para cada materia m_j (4.16) que un estudiante st (4.3) es probable que tome en el semestre j (4.15) dadas sus inscripciones en k (4.4) semestres anteriores $s_a = m_{(j-1)}, \ldots, m_{(j-k)}$ (4.17) en el modelo por omisión en cadenas de Markov se calcula como se muestra a continuación:

$$r(st, m_j, j) = \sum_{s_a = \{m_{(j-1)}, \ldots, m_{(j-k)}\}} W(st, s_a)\, p(s_a \rightarrow s_a \cup \{m_j\}) \qquad \textbf{(4.2.1.27)}$$

Donde $W(st, s_a)$ (4.22) es el peso del estado $m_{(j-1)}, \ldots, m_{(j-k)}$ (4.17) para el estudiante st (4.3). Se utiliza el modelo básico de Markov si no hay semestres omitidos para la construcción de los estados, esto indica que para todos los estados se asigna un peso igual a 1.

4.3. Pseudocódigos

4.3.1 Pseudocódigo para modelo básico de Markov.

Los siguientes pasos describen la implementación basados en el modelo por omisión en cadenas de Markov [7], para predecir las materias que cada estudiante es más probable que tome en su cuarto semestre:

1 **Inicia** /*Algoritmo de Recomendación con Modelo básico de Marcov */
2 **Construir** Cadenas de estados
3 **Buscar** Transiciones de materias consecutivas
4 **Calcular** Puntajes de recomendación
5 **Termina**

En el punto 2 **s**e construyen dos cadenas por estudiante:

1) k cadenas: $\{m_1 \rightarrow m_2 \rightarrow m_3 \rightarrow \ldots \rightarrow m_k\}$ donde m_2 es tomado después de m_1 , m_3 es tomado después de m_2 y así sucesivamente.

2) k+1 cadenas: $\{m_1 \rightarrow \ldots \rightarrow m_k \rightarrow m_{k+1}\}$ donde m_{k+1} es tomado después de m_k .

Para simplificar se asume que una historia de dos semestres consecutivos $(k=2)$ es considerado en el modelo de Markov para el conjunto de datos de la Tabla II. Esto asume que las materias que un estudiante matricula en cada semestre dependen de las materias que tomó en los dos semestres anteriores.

Tabla III. Cadenas de dos y tres materias consecutivas matriculadas por un estudiante

Estudiante	**Cadenas de dos materias**	**Cadenas de tres materias**
st_1 : 1255820000280	BA172→BA372	BA172→BA372→BUB1
	BA172→CB234	BA172→BA372→IS105
	CB115→ BA372	BA172→CB234→BUB1
	CB115→CB234	BA172→CB234→IS105
	BA372→BUB1	CB115→ BA372→BUB1
	BA372→IS105	CB115→ BA372→IS105
	CB234→BUB1	CB115→CB234→BUB1
	CB234→IS105	CB115→CB234→IS105

Nota: Cadenas de dos y tres materias consecutivas para cada estudiante.

Tabla III. (Continuación)

Estudiante	Cadenas de dos materias	Cadenas de tres materias
st_2 : 1259498737788	BA172→BA372	BA172→BA372→CB314
	BA172→CB234	BA172→BA372→CB334
	BA172→CB242	BA172→CB234→CB314
	IS105→BA372	BA172→CB234→CB334
	IS105→CB234	BA172→CB242→CB314
	IS105→CB242	BA172→CB242→CB334
	BA372→CB314	IS105→BA372→CB314
	BA372→CB334	IS105→BA372→CB334
	CB234→CB314	IS105→CB234→CB314
	CB234→CB334	IS105→CB234→CB334
	CB242→CB314	IS105→CB242→CB314
	CB242→CB334	IS105→CB242→CB334
st_3 : 1267745842136	IS105→BA172	IS105→BA172→CB234
	IS105→CB115	IS105→BA172→BA372
	IS142→BA172	IS105→CB115→CB234
	IS142→CB115	IS105→CB115→BA372
	BA172→CB234	IS142→BA172→CB234
	BA172→BA372	IS142→BA172→BA372
	CB115→CB234	IS142→CB115→CB234
	CB115→BA372	IS142→CB115→BA372

Tabla III. (Continuación)

Estudiante	Cadenas de dos materias	Cadenas de tres materias
st_4 : 1278685878516	BA172→BA372	BA172→BA372→CB215
	BA172→IS284	BA172→BA372→CB234
	CB115→BA372	BA172→IS284→CB215
	CB115→IS284	BA172→IS284→CB234
	BA372→CB215	CB115→BA372→CB215
	BA372→CB234	CB115→BA372→CB234
	IS284→CB215	CB115→IS284→CB215
	IS284→CB234	CB115→IS284→CB234

En el punto 3 del algoritmo por omisión, se hacen recomendaciones a un estudiante, es necesario tomar el conjunto de k materias consecutivas tomadas por el estudiante en k semestres y para cada conjunto m , se buscan todas las cadenas $k+1$ en el conjunto de datos iniciando con m . Este paso construye los estados de transición para identificar las siguientes materias que un estudiante podría tomar basado en las materias que ha tomado en k semestres anteriores.

Tabla IV. Todas las cadenas de tres cursos consecutivos en el conjunto de datos iniciando con $\{m_2, m_3\}$

Estudiante	Cadenas de dos materias	Cadenas de tres materias
st_1 : 1255820000280	{BA372, BUB1}	-
	{BA372, IS105}	{BA372, IS105}→CB314
		{BA372, IS105}→CB334
	{CB234, BUB1}	-
	{CB234, IS105}	{CB234, IS105}→CB314
		{CB234, IS105}→CB334
st_2 : 1259498737788	{BA372, CB314}	-
	{BA372, CB334}	-
	{CB234, CB314}	-
	{CB234, CB334}	-
	{CB242, CB314}	-
	{CB242, CB334}	-

Nota: Muestra el resultado del punto 3 para cada estudiante en el conjunto de datos de ejemplo. Se lista un conjunto de dos materias consecutivas tomadas por cada estudiante en semestre del 2 al 3 $m_2 \rightarrow m_3$ y se toman todas las transiciones desde las dos materias consecutivas en todo el conjunto de datos; es decir, todas las cadenas de tres materias consecutivas en el conjunto de datos iniciando con m_2, m_3 .

Tabla IV. (Continuación)

Estudiante	Cadenas de dos materias	Cadenas de tres materias
st_3 : 1267745842136	{BA172, CB234}	{BA172, CB234}→BUB1
		{BA172, CB234}→IS105
		{BA172, CB234}→CB314
		{BA172, CB234}→CB334
	{BA172, BA372}	{BA172, BA372}→BUB1
		{BA172, BA372}→IS105
		{BA172, BA372}→CB314
		{BA172, BA372}→CB334
		{BA172, BA372}→CB215
		{BA172, BA372}→CB234
	{CB115, CB234}	{CB115, CB234}→BUB1
		{CB115, CB234}→IS105
	{CB115, BA372}	{CB115, BA372}→BUB1
		{CB115, BA372}→IS105
		{CB115, BA372}→CB215
		{CB115, BA372}→CB234
st_4 : 1278685878516	{BA372, CB215}	-
	{BA372, CB234}	-
	{IS284, CB215}	-
	{IS284, CB234}	-

En el punto 4 se calcula el puntaje recomendación para cada materia que un estudiante es probable que tome en el siguiente semestre utilizando la formulas (4.2.1.24) y (4.2.1.25).

Para el conjunto de datos se tomó el estudiante st_1 para calcular la probabilidad de que se matricule en la materia CB314 en el cuarto semestre, para esto hay que encontrar las transiciones que conducen a CB314 y calcular sus probabilidades usando MLE (4.2.1.24). Hay dos posibles transiciones para el estudiante st_1 :

1) {BA372, IS105}→{BA372, IS105, CB314}, y

2) {CB234, IS105}→{CB234, IS105, CB314}

La probabilidad de la primera transición es 1/1, la cual es igual a el número de estudiantes que tomaron CB314 después de haber tomado {BA372, IS105} dividido el número de estudiantes que tomaron {BA372, IS105} en dos semestres consecutivos. Algo semejante ocurre con la probabilidad de la segunda transición es igual a 1/1, el cual es igual a el número de estudiantes que matricularon CB314 después de {CB234, IS105} dividido el número de estudiantes que tomaron {CB234, IS105} en dos semestres consecutivos. Y así sucesivamente, el puntaje de recomendación para el estudiante st_1 matriculando CB314 en su cuarto semestre es:

$$r(st_1, CB314) = \frac{1}{1} + \frac{1}{1} = 2 \qquad \textbf{(4.2.1.28)}$$

Las materias que ya han sido matriculadas en semestres anteriores por cada estudiante son excluidas. Para el estudiante st_3 se excluyen las materias recomendadas IS105 y CB234.

Tabla V: Materias recomendadas para cada estudiante

Estudiante	Materias recomendadas para matricular en el siguiente semestre	Puntaje de recomendación
st_1 : 1255820000280	CB314 CB334	2 $1/1+1/1=2$
st_2 : 1259498737788	No hay recomendaciones	-
st_3 : 1267745842136	BUB1 CB314 CB334 CB215	$1/3+1/4+1/2+1/3=1.42$ $1/3+1/4=0.58$ $1/3+1/4=0.58$ $1/4+1/3=0.58$
st_4 : 1278685878516	No hay recomendaciones	-

Nota: En la tabla V se muestran las materias que cada estudiante debería matricular en el siguiente semestre junto con su puntaje de recomendación.

Se debe tener presente que cuando se realiza el calculo con todas las materias del estudiante el valor del puntaje recomendado puede variar por haber más conjuntos de materias como se muestra en el Anexo E.

4.3.2 Pseudocódigo para modelo por omisión en cadenas de Markov.

Para algunos de los estudiantes de nuestro conjunto de datos de ejemplo no se encontró alguna recomendación de materias que debería tomar en el siguiente semestre, esta casuística se da cuando el conjunto de datos de materias consecutivas matriculadas por un estudiante en k

semestres anteriores no coincide con la de otros estudiantes. En la tabla V se presentan los estudiantes st_2 y st_4 con este mismo tipo de casuística, en el cual no hay recomendación de materias que debería matricular en el cuarto semestre, esto ocurre porque las materias matriculadas en los semestres dos y tres no coinciden con ninguna secuencia de cadenas de materias en el conjunto de datos.

Los siguientes pasos describen la implementación del modelo por omisión en cadenas de Markov basados en la investigación [7], para predecir las materias que cada estudiante es más probable que tome en su cuarto semestre:

1 **Inicia** /*Algoritmo de Recomendación con Modelo por omisión en cadenas de Marcov */
2 **Buscar** Estudiantes que no tienen materias recomendadas
3 **Buscar** Transiciones de materias con cadenas omitidas
4 **Calcular** Puntajes de recomendación
5 Termina

En el punto 2 se obtienen los estudiantes que no tienen recomendación de materias o que tienen menos de seis materias recomendadas $CountR_{m_{(k+1)}}$ (4.12) para matricular en el siguiente semestre $k+1$ (4.5).

Se seleccionó el estudiante st_4 de la tabla II, por no tener ninguna materia recomendada, y se aplica el modelo por omisión en cadenas de Markov, en el cual se construye las cadenas de materias omitiendo el segundo semestre, es decir, se omite un solo semestre λ^1 (4.21).

Tabla VI: Cadenas de dos materias omitiendo el segundo semestre

Estudiante	Cadenas de dos materias
st_4 : 1278685878516	BA172→CB215
	BA172→CB234
	CB115→CB215
	CB115→CB234

Nota: Cadenas de dos materias omitiendo el segundo semestre para el estudiante st_4 .

Tabla VII: Todas las cadenas omitidas de tres cursos consecutivos en el conjunto de datos iniciando con $\{m_2, m_3\}$

Estudiante	Cadenas de dos materias	Cadenas de tres materias
st_4 : 1278685878516	BA172→CB215	-
	BA172→CB234	{BA172, CB234}→BUB1
		{BA172, CB234}→IS105
		{BA172, CB234}→CB314
		{BA172, CB234}→CB334
	CB115→CB215	-
	CB115→CB234	{CB115, CB234}→BUB1
		{CB115, CB234}→IS105

Nota: Para el punto 3 del algoritmo por omisión, se encuentran las cadenas con semestres omitidos, en la cual algunas cadenas de materias de dos cursos coinciden con otras transiciones de materias de otros estudiantes.

En el punto 4 del modelo por omisión, se realiza el calculo del puntaje de recomendación para las materias que aparecen en la columna 3 de la tabla VII y se asume que $\lambda = 1/2$:

$$r(st_4, BUB1) = W(st_4, \{BA172, CB234\}) * P(\{BA172, CB234\} \rightarrow BUB1) + W(st_4, \{CB115, CB234\}) * P(\{CB115, CB234\} \rightarrow BUB1) \quad \textbf{(4.2.1.28)}$$

$$r(st_4, BUB1) = (\frac{1}{2} * \frac{1}{3}) + (\frac{1}{2} * \frac{1}{2}) = 0.41 \quad \textbf{(4.2.1.29)}$$

Tabla VIII: Materias recomendadas para el estudiante st_4 , en el modelo por omisión de cadenas de Markov

Estudiante	**Materias recomendadas para matricular en el siguiente semestre**	**Puntaje de recomendación**
st_4 : 1278685878516	BUB1	0.41
	IS105	0.41
	CB314	0.17
	CB334	0.17

Nota: Se encuentra las materias recomendadas para el estudiante st_4 con su respectivo puntaje de recomendación.

4.4. Resultados obtenidos

Se implementó un análisis para los cuatro estudiantes del conjunto de prueba en el que se calculó el Recall como, el porcentaje de materias que se recomendaron por el sistema para matricular

para el siguiente semestre y que ya fueron tomadas por los estudiantes en sus semestres anteriores. Por otra parte, se calculó la precisión como, el porcentaje de materias que fueron matriculadas en el siguiente semestre basado en los datos de prueba.

El cálculo del Recall y la Precisión realizado para los cuatro estudiantes de prueba, en el modelo básico o por omisión este modelo puede predecir el 68,75 y 80,77 porciento respectivamente, de las materias que el estudiante ya matriculó en semestres anteriores; teniendo en cuenta que estas materias se eliminan del resultado final que se le muestra al estudiante para que matricule en el próximo semestre. Por otra parte, el cálculo de la precisión indica que el 28,57 porciento de las materias recomendadas a los estudiantes fueron tomadas por estos en el siguiente semestre según los datos de prueba para el modelo Básico de Markov y se aumentó este valor con respecto al primer modelo utilizado en 14,29 porciento utilizando el modelo por omisión en cadenas de Markov, por ello se está acertando no solo en las materias recomendadas sino que se está teniendo en cuenta el proceso que los estudiantes tienen durante todos los semestres anteriores.

Tabla IX: Rendimiento de modelo básico y por omisión en Cadenas de Marcov

Modelo	**Recall**	**Precisión**
Modelo básico de Marcov	68,75%	28,57%
Modelo por omisión en cadenas de Markov	80,77%	42,86%

Nota: cálculo del Recall y la Precisión realizado para los cuatro estudiantes de prueba, en el modelo básico o por omisión.

Es necesario dejar claro que la calidad de los modelos nos permite tener la precisión suficiente para confiar en el sistema de recomendación automático, en el cual se encuentra que la precisión de las materias que debe matricular el estudiante en el siguiente semestre es del 28,57 porciento y que el 68,75 porciento de las materias recomendadas no apoyan a la calidad de los datos para el modelo básico de Markov; y para el Modelo por omisión la precisión de las materias que debe matricular el estudiante en el siguiente semestre es del 42,86 porciento y que el 80,77 porciento de las materias recomendadas no apoyan a la calidad de los datos.

Además de estos datos se debe tener presente que el sistema recomienda materias que el estudiante no ha matriculado en semestres anteriores pero que en los datos de pruebas estas materias no aparecen matriculadas por estudiantes en el semestre evaluado.

4.5. Python

En la inteligencia artificial existe la necesidad de construir estructuras de datos y algoritmos que requieren de una gran capacidad de recursos de la máquina, por lo tanto es importante considerar la construcción de algoritmos que permitan tomar un camino hacia el rendimiento en la comunicación de los artefactos, para ello se utilizó un lenguaje de programación de alto nivel llamado Python.

El lenguaje de programación Python fue originalmente desarrollado por Guido Van Rossum cerca de 1990 [18], y desde entonces se ha convertido en un lenguaje muy utilizado en la industria y la educación. La segunda mayor versión de este lenguaje, Python 2, fue publicada en 2000, y la tercera mayor versión, Python 3, publicada en 2008. Hay considerables diferencias e incompatibilidades entre Python 2 y Python 3. Para esta investigación se utilizó la tercera versión, Python 3, específicamente, Python 3.6.8.

Python es un lenguaje interpretado, sus comandos son ejecutados a través de un software conocido como el intérprete de Python. El intérprete recibe un comando, evalúa ese comando e informa el resultado del comando. Mientras que el intérprete se puede usar de forma interactiva, especialmente cuando se depura, un desarrollador normalmente define una serie de comandos y los guarda en un archivo de texto plano conocido como código fuente o script. Para Python, el código fuente se guarda convencionalmente en un archivo con formato ".py" [38].

5 CAPITULO V: REFERENCIA BIBLIOGRÁFICA, RECOMENDACIONES Y CONCLUSIONES

5.1. Conclusiones y trabajos futuros

Para este proyecto se tuvo acceso a la información procesada en la que se accedió a todas las materias que matricularon los estudiantes durante el proceso de su carrera universitaria hasta su graduación. Por medio del modelo implementado se demostró que la secuencia de las materias tiene gran importancia en la recomendación para la matrícula de nuevas materias en los siguientes semestres.

Se considera este proyecto como un piloto que permite preparar los datos de los estudiantes para ser analizados por medio de un modelo que recomienda varias materias a los estudiantes para los siguientes semestres. Por lo tanto, Al no utilizar ningún conocimiento previo institucional para construir sistemas de recomendación de materias que debe matricular un estudiante en el siguiente semestre, se está construyendo un modelo que puede ser mejorado para que el porcentaje de precisión de recomendación tenga una mejor calidad.

Este proyecto busca apoyar a los estudiantes a seleccionar mejores materias para matricular en el siguiente semestre, teniendo en cuenta la información de estudiantes que ya pasaron por este proceso y que se graduaron con éxito. Este modelo tiene el conocimiento pleno de los procedimientos realizados por anteriores estudiantes, por lo que a la hora de matricular materias el modelo puede ser mucho más efectivo que si lo hubiere hecho el estudiante sin ninguna ayuda automática.

Se recomienda para trabajos futuros la creación de micro-servicios que permitan utilizar diferentes modelos matemáticos según sea la necesidad como recomendación, clasificación u otro método. En estos micro-servicios se permite enviar parámetros de las materias vista en cada uno de los semestres con un formato específico que pueda ser entendido por el sistema y finalmente este pueda procesar la información y dar un resultado de las materias recomendadas a matricular en el siguiente semestre.

Además de lo anterior, la precisión se puede aumentar significativamente si se utilizan conocimientos académicos adicionales para procesar y mejorar las recomendaciones generadas por el sistema, como los requisitos previos de matrícula, las materias básicas y electivas. Por otra parte se puede adicionar información socio-demográfica para realizar una clasificación de los estudiantes que tienen características similares y finalmente tener información mucho más exacta para realizar la recomendación de las materias.

5.2. Bibliografía

[1] C. Raigosa Hernández and J. E. Pabon Betancurt, "Nivel de deserción universitaria en el programa de ingeniería de sistemas y computación de la universidad tecnológica de Pereira", Universidad Tecnológica de Pereira, 2017.

[2] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, "Predicting at-risk university students in a virtual learning environment via a machine learning algorithm", *Comput. Human Behav.*, Jun. 2018.

[3] T. W. Anderson and L. A. Goodman, "Statistical Inference about Markov Chains", *Ann. Math. Stat.*, vol. 28, no. 1, pp. 89–110, Mar. 1957.

[4] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey", 2015.

[5] M. Kandlhofer, G. Steinbauer, S. Hirschmugl-Gaisch, and P. Huber, "Artificial intelligence and computer science in education: From kindergarten to university", in *2016 IEEE Frontiers in Education Conference (FIE)*, 2016, pp. 1–9.

[6] C. E. Lopez Guarin, E. L. Guzman, and F. A. Gonzalez, "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining", *IEEE Rev. Iberoam. Tecnol. del Aprendiz.*, vol. 10, no. 3, pp. 119–125, Aug. 2015.

[7] E. S. Khorasani, Z. Zhenge, and J. Champaign, "A Markov chain collaborative filtering model for course enrollment recommendations", in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 3484–3490.

[8] ODES (Observatorio de Eeducación Superior de Medellín), "Deserción en la educación superior", Medellin, 2017.

[9] M. Marta Ferreyra, C. Avitabile, J. Botero Álvarez, F. Haimovich Paz, and S. Urzúa, "At a Crossroads Higher Education in Latin America and the Caribbean Human Development.", 2017

[10] D. Pamela Brito Fuentes, Monica Cristina Pineda Arroyo, Monica Ospina Londoño, "Análisis de la deserción y la graduación de los estudiantes de la Universidad de la Guajira", 2016.

[11] A. F. Ramirez Correo and T. M. David, “Causas de la deserción en el programa I.S.C de la Universidad Tecnologica de Pereira”, 2017.

[12] Universidad Tecnológica de Pereira, “Oracle BI Publisher.” [Online]. Available: http://reportes.utp.edu.co/xmlpserver/publico/Planeacion/Academica/Desercion/DIA/DIA_periodo.xdo?_xmode=2. [Accessed: 02-Aug-2019].

[13] Pereira como vamos - Programa de seguimiento y evaluación de la calidad de vida en Pereira, “Informe de calidad de vida en Pereira - ICV 2015”, 2015.

[14] Cámara de Comercio de Pereira, “Educación superior y la econimía de Pereira.” [Online]. Available: https://www.camarapereira.org.co/es/ieventos/ver/2328/educacion-superior-y-la-economia-de-pereira/. [Accessed: 25-Jul-2019].

[15] Universidad Tecnológica de Pereira, “Oracle BI Publisher.” [Online]. Available: http://reportes.utp.edu.co/xmlpserver/publico/Planeacion/Boletin_estadistico/Inscritos/inscrito.xdo;jsessionid=MhKndGKTyqxjmqRhQGy4k2l8Td0814vMJBXVy5MvCGxmsdQk7sxT!-1553369420?_xmode=2. [Accessed: 02-Aug-2019].

[16] Universidad Tecnológica de Pereira, “Programa de ingeniería de sistemas y computación,” Pereira, 2019.

[17] C. Fernández Collado and L. Pilar Baptista, “Metodología de la investigación”, vol. 6. Mexico, 2014.

[18] M. T. Goodrich, R. Tamassia, and M. H. Goldwasser, “Data structures and algorithms in Python”, 2013.

[19] N. J. Nilsson, “Logic and artificial intelligence”, *Artif. Intell.*, vol. 47, no. 1–3, pp. 31–56, 1991.

[20] W. Sun and X. Gao, “The Construction of Undergraduate Machine Learning Course in the Artificial Intelligence Era”, in *2018 13th International Conference on Computer Science & Education (ICCSE)*, 2018, pp. 1–5.

[21] B. Fester, “Teaching Machines (Ch 6): Making Sense of Teaching Machines”, in *Teaching Machines*, Baltimore: Johns Hopkins University Press, 2014.

[22] P. J. Hayes, K. M. Ford, and J. R. Adams-Webber, “Human Reasoning about Artificial Intelligence”, in *Thinking Computers and Virtual Persons*, Elsevier, 1994, pp. 331–353.

[23] L. Sijing and W. Lan, “Artificial Intelligence Education Ethical Problems and Solutions”, in *2018 13th International Conference on Computer Science & Education (ICCSE)*, 2018, pp. 1–5.

[24] M. E. Álvarez and R. Orrego, “Modelo Markoviano para el estudio de evolución de cohortes de estudiantes de un programa académico”, *Rev. Univ. EAFIT*, vol. 36, no. 120, pp. 45–56, 2000.

[25] M. Ciolacu, A. F. Tehrani, L. Binder, and P. M. Svasta, “Education 4.0 - Artificial Intelligence Assisted Higher Education: Early recognition System with Machine Learning to support Students’ Success”, in *2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, 2018, pp. 23–30.

[26] S. Helal *et al.*, “1Predicting academic performance by considering student heterogeneity”, *Knowledge-Based Syst.*, vol. 161, pp. 134–146, Dec. 2018.

[27] C. A. Javier and M. M. C. Ariel, “Factores determinantes de la deserción.”, Spadies, Agosto 2016

[28] D. oficial de la Federación, “Diseño de un modelo estocástico usando cadenas de Markov para pronosticar la deserción académica de estudiantes de ingeniería. Caso: Escuela Colombiana de Ingeniería Julio Garavito”, vol. 10, no. 9, p. 32, 2012.

[29] E. Caro, C. González, and J. M. Mira, “Student academic performance stochastic simulator based on the Monte Carlo method”, 2014.

[30] P. Vuttipittayamongkol, “Predicting factors of academic performance”, in *2016 Second Asian Conference on Defence Technology (ACDT)*, 2016, pp. 161–166.

[31] Q. A. Al-radaideh, “Mining Student Data Using Decision Trees”, no. January 2014.

[32] I. Krikun, “Applying learning analytics methods to enhance learning quality and effectiveness in virtual learning environments”, in *2017 5th IEEE Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, 2017, pp. 1–6.

[33] M. F. Uddin and J. Lee, “Predicting good fit students by correlating relevant personality traits with academic/career data”, in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 968–975.

[34] R. Otero, S. Bolívar, and J. Palacios, “Análisis de la retención de estudiantes de ingeniería basado en la pérdida consecutiva de una misma asignatura. Un enfoque de Cadenas de Markov Retention analysis of engineering students based on consecutive course failure. A

Markov Chain Approach", *Ing. Ind. Actual. y Nuevas Tendencias*, vol. 5, no. 16, pp. 7–18, 2016.

[35] P. Kumar and R. S. Thakur, "Recommendation system techniques and related issues: a survey", *Int. J. Inf. Technol.*, vol. 10, no. 4, pp. 495–501, Dec. 2018.

[36] M. Lobur, M. Shvarts, Y. Stekh, and I. Demkiv, "Some methods for predicting recommendations for MEMS designer communities", in *2018 14th International Conference on Perspective Technologies and Methods in MEMS Design, MEMSTECH 2018 - Proceedings*, 2018, pp. 196–199.

[37] M. Khalaji and N. Mohammadnejad, "CUPCF: combining users preferences in collaborative filtering for better recommendation", *SN Appl. Sci.*, vol. 1, no. 9, pp. 1–9, Sep. 2019.

[38] "Welcome to Python.org." [Online]. Available: https://www.python.org/. [Accessed: 09-May-2020].